Argumentation for Statistical Model Selection

Sassoon, Isabel Karen

Awarding institution:
King's College London

Argumentation for Statistical Model Selection

by

Isabel Karen Sassoon

A thesis submitted in partial fulfillment for the degree of Doctor of Philosophy

in the
Department of Informatics
of the
Faculty of Natural & Mathematical Sciences

December 2017

To Yair, Tom and Emma

Abstract

The increased availability of clinical data, in particular case data collected routinely, provides a valuable opportunity for analysis with a view to support evidence based decision making. In order to confidently leverage this data in support of decision making, it is essential to analyse it with rigour by employing the most appropriate statistical method. It can be difficult for a clinician to choose the appropriate statistical method and indeed the choice is not always straight forward, even for a statistician. The considerations as to what model to use depend on the research question, data and at times background information from the clinician, and will vary from model to model.

This thesis develops an intelligent decision support method that supports the clinician by recommending the most appropriate statistical model approach given the research question and the available data.

The main contributions of this thesis are: identification of the requirements from real-world collaboration with clinicians; development of an argumentation based approach to recommend statistical models based on a research question and data features; an argumentation scheme for proposing possible models; a statistical knowledge base designed to support the argumentation scheme, critical questions and preferences; a method of reasoning with the generated arguments and preference arguments. The approach is evaluated through case studies and a prototype.

Acknowledgements

Firstly, I would like to thank Professor Mark McGurk for setting up the studentship that funded this research. Mark's passion for the collection and the exploitation of clinical data was an essential catalyst to this research, which literally could not have happened were it not for Mark. The data he has meticulously collected and provided to this project has ensured that this was all delivered using real world examples. He was also instrumental in setting up collaborations with clinicians which will continue to shape my research beyond the scope of this thesis.

Secondly, I would like to thank my supervisors Jeroen Keppens and Peter McBurney for their constant support, encouragement and patience. I have thoroughly enjoyed working with them and will miss engaging in stimulating discussions with them on a regular basis. Their timely, meticulous and constant feedback on my work has been crucial to my progress.

Thirdly, there are many other people across the department of informatics who have provided helpful advice, guidance and mentoring. I am particularly grateful to Sophia Tsoka for taking on the role of my external examiner through the internal review sessions, to Elizabeth Sklar and Simon Parsons for coaching and mentoring, and to the members of the Agents and Intelligent Systems research group for providing feedback and perspective on my work. I am grateful to Sebastian Zillesen for taking on the development of the prototype as part of his MSc. project, and delivering above and beyond the scope of the project.

A special thank you to my examiners Professor Francesca Toni and Dr Floris Bex for their constructive feedback and detailed comments on this thesis.

During my time as a PhD student I submitted various papers to conferences, workshops and journals. Independently of the result, I valued the anonymous reviewers who took the time to review and provide feedback and in some cases point me in some extremely interesting directions. Even when the result was not favourable I was able to use the feedback to further my research.

My final thanks must go to my family and my children for their ongoing support and encouragement. Needless to say that without my husband's support, encouragement and the bespoke desk he built for me this thesis would not have been possible.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

"Data without a model is just noise"

- The End of Theory: The Data Deluge Makes the Scientific Method Obsolete by C. Anderson, *Wired 2008* [8]

In this chapter I outline my research question and provide an overview of the topics relevant to the research. I also include an overview of the structure of this thesis.

1.1 Background

The collection of data as routine is becoming common practice in all fields, in the context of clinical data this has generated a wealth of data sources. These sets of data will be collected under different circumstances, some will be administrative and collected by the hospital IT systems, whilst some will be collected directly from the patient by the individual departments.

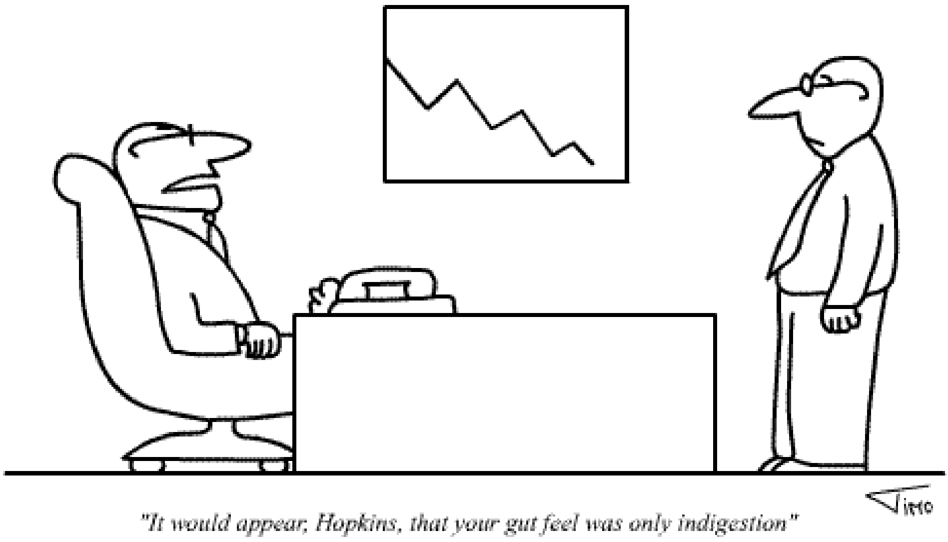


FIGURE 1.1: "It would appear Hopkins, that your gut feel was only indigestion" Data driven vs. gut feel decisions

There is a need to exploit this data more routinely as part of evidence based decision making to inform clinical practice. In order to do so there is a requirement for clinicians to be able to access this data, but more importantly for clinicians to be able to either explore the data or use it to test their hypotheses. In addition to being able to do the latter it is also important that this can be done without the need of statistical, informatics or administrative support. This will reduce the time it takes to obtain robust answers to data based research question by removing the necessity to involve additional resources, which may be scarce.

In order to leverage this data in support of evidence based decision making it is essential that the data is analysed with rigour. This rigour is based on using the most appropriate statistical method in the context of the analysis objective, the hypothesis to be tested and the data. The use of the most appropriate method will provide confidence in any conclusions derived from the analysis of the data.

For the clinician tasked with exploiting this data the optimal choice of model is not always straight forward and in some cases the choice is not clear cut. The considerations as to what model to use depend not just on the clinician's research question and data but may also depend on background information from the clinician, and may vary from model to model. Two models suitable to achieve the same analysis objective may perform better under different circumstances. For example some models are more robust when there is a high proportion of missing data, so this consideration will be relevant only under those circumstances. Easy to use statistical software packages make the analysis easy but offer no guidance on selecting the most appropriate model given the circumstances.

It is also important that any recommendation on the most appropriate statistical model to use in each case includes the reasoning behind the recommendation. There will be situations where more than one model is appropriate and therefore the merits of each approach will need to be noted, and opinions may vary as to what is the 'best' method to use. In addition to more than one model being appropriate the clinician's preferences or some features present in the data or related to the research question may result in one approach being more suitable than another.

Some of the information required in order to select the best statistical approach may be incomplete or missing. As the data is collected as routine there may be situations where as the data grows the approach to testing the same research question or hypothesis will be best answered with different techniques. This can be caused by the change in the overall data and how it would lend itself to being analysed by different methods. For example, some statistical methods are more suitable for small samples, whilst others will become options only once the volume of data grows beyond a specific critical number of cases or rows.

1.1.1 Selecting the statistical model

"All models are wrong but some are useful"

- George Box, *[18]*

When looking to test a research question or hypothesis clinicians would interact with statistical concepts at the design stage of a study, when selecting the models to use to analyse the data, and when performing and interpreting the analysis. This thesis focuses on the middle part of the process, the selection of the model because I am focusing on exploitation of existing data rather than data collection strategies.

Clinicians may not always be qualified in performing the analysis required in support of their research question and as such would involve a statistician. The statistician's role it is to understand the data in the context of the research questions and to recommend the statistical analysis approach or model best suited to provide the results required. The need to consult with a statistician for these analyses can be a blocker or a barrier to the analysis being performed. The use of a model that is not appropriate can result in rejection or revisions when submitting the result of an analysis to a journal.

> *"To consult the statistician after an experiment is finished is often merely to ask him to conduct a post mortem examination. He can perhaps say what the experiment died of."*

- R. Fisher, *[35]*

In practice in order for clinicians to answer a specific research question on existing data, the first step of the analysis involves the identification of the analysis objective. The analysis objective depends primarily on the target variable of interest (dependent variable). Some examples of types of target variables are: nominal, interval or time to event. For each of these target variable types there can be multiple possible statistical model approaches, these approaches differ in the assumptions they make as well as the circumstances under which they are most effective.

An example of a clinician's research question could be *Is there a difference in survival time between patients on treatment A and patients in treatment B?* In this case the target variable of interest is the survival time, and this corresponds to an analysis objective of time to event. Another example could be *Is there a difference in BMI (Body Mass Index) between patients on treatment A and treatment B?* In this case the target variable of interest is the BMI, which is a numerical interval variable and this corresponds to an analysis objective of Interval.

Often, in empirical analyses of clinical data, models are chosen poorly or cannot be justified. A systematic review by Abraira *et al.* [1] highlighted that reporting of survival analysis results, one important type of empirical analysis of clinical data, had increased within journal publications; however the quality of the reporting of the statistical analysis was improving slowly. More pertinent to the aim of this thesis is that Abraira *et al.* found a low proportion of articles that mention validation of model assumptions prior to use (proportional hazards assumptions for Cox modelling [24] as a specific example).

1.1.2 Clinical Analysis requirements

The analysis and hypothesis testing on a data set often involves more than one clinician and multiple different subsegments. This results in different research questions being applied to the same data set, and as such there is a need to provide a justification for any model recommendation made. Each individual clinician will provide different preferences and will assess the assumptions differently, as such in some situations the same data set and research question can yield a different recommended model. The clinical setting reinforces the need for some fundamental requirements from any method employed in statistical model selection. These are an ability to deal with conflicting conclusions, an ability to handle incomplete information, and the facility to provide justification for the resulting recommendation.

1.2 The Thesis Research Question

The research in this thesis focuses on delivering a methodology that supports clinicians during the task of answering research questions through testing hypotheses on existing data. Clinicians are assumed to be the target end user to leverage this methodology within the scope of this thesis. The proposed approach enables more independent and robust analysis by proposing an intelligent model selection system to address the challenges of statistical model selection. The model selection methodology proposed in this thesis suggests the appropriate model(s) to the clinician based on the research question, the data and any external relevant input.

In order to deliver a methodology to support the process of statistical model selection in this context there are various challenges that need to be addressed. Firstly, one of the foundations of the methodology is a representation of the statistical theory. The latter dictates what statistical models are relevant to answer different types of research questions. The representation of this knowledge also has to support the validation process for the suitability of each model. This can be assessed through assumption testing. The latter needs to fulfil the assumption tests by either leveraging the available data or asking the clinician or end user. Furthermore the process needs to be transparent as the confidence in any recommendation made will depend on the inference process being retained and explained to the clinician.

Situations arise when more than one model is suitable to be applied to a given data set and research question. The option of applying all the possible models may not be appropriate, as the situation can arise where each of multiple suitable models will generate different results or conclusions. Therefore a crucial aspect this thesis aims to address is the generation of a set of recommended models that take into account any preferences that make some possible models more suitable than other possible models, hence recommended. These preferences are derived from both statistical theory and from the clinician's preferences for a model. There is a need to balance the importance of the different types of preferences within the process, and to provide an audit trail of the preferences applied to each recommendation made. This makes it possible to compare the recommended analysis approach on a given set of data, under different clinician preferences.

The approach that I have taken to develop the methodology has been based on argumentation, as the process of recommending one analysis approach over another involves weighing up the relative pros and cons of the different options. Furthermore argumentation has been shown through its use in decision support to accommodate for requirements such as those mentioned in Section 1.1. An additional desired feature is the ability to be driven by a separate knowledge base, to make the system flexible and expandable by altering the knowledge base rather than the process.

> The aim of this thesis is to develop an intelligent model selection system that recommends the most appropriate model to the clinician, based on their research question, the data, and any relevant external input. The methodology needs to take input from both the data and the clinician, to represent statistical knowledge, to reason with conflicting preferences and to justify the recommendation made.

The approach I have taken also aims for a methodology that can be applied to any domain or data. The case studies covered and requirements considered in this thesis have been clinical in nature but this work can and will be applied to different domains in the future. Given the nature of the data and research questions considered then it is assumed in this thesis that the end user is a clinician.

1.3 Case Studies

The research question addressed in this thesis is a result of the collaboration with the maxillofacial oncology department at Guy's hospital. The clinicians in the department are routinely collecting and have access to a broad range of clinical data. The department is striving to collect data directly from patients as routine at each patient visit through the use of hand held devices.

These data sources are an asset when clinicians wish to test hypotheses and gain further understanding of patients and their progression. I will introduce two examples of the type of data analysis projects and objectives. These examples will be used as case studies in later chapters to assist with the evaluation of the original contributions I present in this thesis. I will also be making use of a separate example of an analysis based on publicly available data as a running example to illustrate the process of statistical model selection as well as illustrating the proposed contributions.

- Sentinel European Node Trial (SENT) is a European multi centre prospective study of the use of sentinel node biopsy. The objective of this study was to establish whether the technique was both reliable in staging the $N0$ neck and a safe oncological procedure in patients with early stage oral squamous cell carcinoma (SCC). The initial findings have been published in [70]. The trial recruited 420 patients from 2005 - 2010 across fourteen European centres and followed their progress and treatment. The main hypothesis or research question of interest for this trial was assessing the difference in diagnostic accuracy between the use of Sentinel Node Biopsy when compared to the standard approach involving a neck dissection. This provided a rich data set for the testing of a wide range of research questions, as well as the main hypothesis this data was collected for. More than 40 clinicians are involved across the centres. The secondary analyses on this data are many and varied and initiated by different clinicians all involved in this trial. Examples of secondary hypotheses considered on this data included testing whether there was a significant difference in survival given tumour or demographic characteristics. This situation would benefit from a method for guiding the selection of the analysis model for each of these secondary analysis, as many clinicians are involved there are also many different approaches and supporting reasons for each analysis. There is also a need for overall consistency and in cases where different analysis approaches result in contradicting conclusions, there is a need to compare the justifications for the approach used in each analysis in order to re-conciliate. Another analysis based on this same database of patients is available in [77].

- Benchmarking Complications - This study looked at developing a model to characterise and assist in predicting the likelihood of patients suffering post operative complications following surgery for the removal of Oral SCC. This data brought together patient records from three different centres. The research question in this case looked at what factors (pre-operative and post-operative) were indicators of higher risk of post-operative complications. The initial work was published [78], and more recently an additional centre of data has been added to the cohort. Two subsequent papers have now been published where the models proposed in [78] are validated against the new data from the additional centre [80], and introduced some additional machine learning approaches to build the models [79]. In this case the relevance of the proposed methodology is justifying the approaches used or any changes in approach in view of the availability of new data, and the justification for applying different or alternative modelling approaches.

1.4 Contributions of this thesis

My research contributions and main goals of the thesis are:

- The role of argumentation in the task of recommending the appropriate statistical model to be used given a research question and available data [68]. (Chapter 3)

  This starts by formulating the problem in a format that is amenable to the application of argumentation. The steps involved include the instantiation of argumentation schemes to create arguments in favour of the use of a particular model, the argument representation and structure of the different model options as arguments, the testing of the assumptions for each of the models available to the specific research objective, and refinement to a list of possible models to be used for the analysis.

- The structure of the argumentation schemes and their associated critical questions. (Chapter 4) This proposes an argument scheme for statistical model selection and articulates appropriate critical questions. The critical questions are instantiated themselves as argumentation schemes and provide rebuttals or undercuts for the arguments instantiated by the argumentation scheme.

- The structure and role of the Statistical Knowledge Base (SKB) in the model selection and argumentation process.(Chapter 3 and 5).

  This focuses on the structure and information required in the SKB in support of the proposed argumentation process. This includes the documentation of the relations between analysis objectives, models and assumptions as well as the method used to validate each assumption. The SKB is then further extended in support of the preferences.

- Incorporating preferences in the model selection process (Chapter 5)

  There are cases where one or more models will be possible and suitable for a given analysis objective and data. In such cases there are other factors such as preferences or circumstances that could be used to recommend the most appropriate model(s). The preferences can incorporate the clinicians' choice for model, the purpose of the analysis (importance of accuracy in predictions) and the contextual factors that make one model more suitable than another one under certain conditions. In this chapter I propose a modification and extension of the methodology that is able to leverage these additional considerations. Part of this work has been published in [69].

- Formalising the original contributions of the thesis (Chapter 6) The original contributions made in this thesis are formalised using Z notation. The formalisation in Z notation provides a set of elements and schemas that facilitated the implementation of a prototype.

- Evaluating the original contributions of the thesis (Chapter 7)

  The original contributions proposed in this thesis are evaluated through their application to case studies. These are taken from analysis introduced in Section 1.3. The evaluation process ensures that the same logic as was applied during the analysis work is reflected within the proposed methodology.

This work is a contribution to automated decision making in artificial intelligence, to computational argumentation and to medical decision making in health informatics.

1.5 Structure of the thesis

This thesis is structured into nine chapters as follows:

Chapter 1 defines the research question, as well as providing an overview of the requirements and motivating case studies.

Chapter 2 presents the literature review of the existing research in areas relevant the main contributions in this thesis.

Chapter 3 provides an overview of the process of statistical model selection, introduces a running example to articulate the elements proposed within this thesis. The approach makes use of argumentation schemes and a knowledge base. This chapter further introduces the structure required of the knowledge base which is the initial original contribution made in this thesis.

Chapter 4 presents the proposed argumentation schemes in support of the model recommendation process as well as their critical questions.

Chapter 5 introduces preferences, defines preferences relevant to the process of statistical model selection and extends the knowledge base to support their use in the process.

Chapter 6 formalises the argument schemes, critical questions and extended statistical knowledge base in Z notation.

Chapter 7 discusses the overall process of evaluation, initially evaluates the original contributions of this thesis through the case studies, describes a prototype implementation and articulates the future evaluation plan.

Chapter 8 provides conclusions on the original contributions proposed in this thesis and articulates future research directions.

Appendix A Z notation

Appendix B Statistical Knowledge Base contents

1.6 Publications

Some aspects of the work described in this thesis has been published in the following:

- The argument scheme and knowledge base structure was initially presented at the *Fifth International Conference on Computational Models of Argument - COMMA 2014* and was published in the proceedings in [68]. The contents of the paper overlap with some of the material presented in chapters 3 and 4.
- The extension of the knowledge base and the implementation of context domains in conjunction with extended argumentation frameworks was presented at the *Sixth International Conference on Computational Models of Argument - COMMA 2016* and was published in the proceedings in [69]. The contents of the paper overlap with some of the material presented in chapter 5.
- My collaborations with clinicians have also resulted in the following publications:
  - The publications related to the analysis work on SENT data were: *Sentinel European Node Trial (SENT): 3-year results of sentinel node biopsy in oral cancer* by Schilling *et al.* published in the *European Journal of Cancer* [70] and *Sentinel Node in Oral Cancer: The Nuclear Medicine Aspects. A Survey from the Sentinel European Node Trial* by Tartaglione *et al.* published in *Clinical Nuclear Medicine* [77].
  - The analysis work related to the Complications data *Is benchmarking possible in audit of early outcomes after operations for head and neck cancer?*

was published by Tighe *et al.* in the *British Journal of Oral and Maxillofacial Surgery* [78]. Additional publications related to the Complications data: *Development of a benchmarking tool for audit of early outcomes after surgery for Head and Neck Squamous Cell Carcinoma* by Tighe *et al.* in the *British Journal of Oral and Maxillofacial Surgery* [79] and *Developing a risk stratification tool for audit of outcome after surgery for head and neck squamous cell carcinoma* by Tighe *et al.* in *Head & Neck* [80].

– The analysis work related to the Implant data *Prophylactic use of pentoxifylline and tocopherol in irradiated head and neck oncology patients requiring dental extractions* [62] and *Use of pentoxifylline and tocopherol in the management of osteoradionecrosis* [63] were both published by Patel *et al.* in the *British Journal of Oral and Maxillofacial Surgery*.

Chapter 2

Literature Review

In this chapter I present a literature review of the areas of research that are relevant to the issues addressed in this thesis. Section 2.1 is a review of the research in the area of expert systems for statistical analysis. Section 2.2 begins a short overview of the areas of argumentation that I will cover. Section 2.3 focuses on the internal structure and generation of an argument. Section 2.4 focuses on the process of reasoning with sets of arguments. Section 2.5 summarises the argumentation aspects covered in previous sections. Section 2.6 reviews applications of argumentation in clinical decision support, including the integration of knowledge bases. Section 2.7 introduces preferences and Section 2.8 provides an overall summary of the literature review.

2.1 Statistical Expert Systems

An overview and introduction on the application of expert systems to the process of data analysis is provided by Hand [43]. In his paper Hand firstly provides a few definitions of an expert systems:

> *"An "Expert System" is regarded as the embodiment within a computer of a knowledge based component, from an expert skill, in such a form that the system can offer intelligent advice or take an intelligent decision about a processing function. A desirable additional characteristic, which many would consider fundamental, is the capability of the system, on demand, to justify its own line of reasoning in a manner directly intelligible to the enquirer. The style adopted to attain these characteristics is rule-based programming"*

- British Computer Society's Committee of the Specialist Group on Expert Systems, *Feb, 1983*

In his paper Hand examines the scope of an expert system for statistical analysis and suggests two modes such a system could employ. The system can be purely a recommendation engine, or it can recommend and perform the recommended analysis. Hand refers to the latter as the oracle mode. He argues that the oracle approach has two major drawbacks: firstly effective statistical work involves interaction between the statistical theory and the domain knowledge, and secondly a statistician would not use this as it leaves no opportunity to understand the recommendation made by the system and therefore does not offer the opportunity to compare the reasoning process to their own.

Another aspect that is relevant to the requirements of an expert system in support of data analysis is the end user. The end user of a statistical expert system can be a novice or an expert and this will affect the desirable properties of such a system. Hand delves into the details of how a statistical consultant goes about selecting the appropriate model. Studies have shown that there are marked similarities in the way statisticians decide which method to use and how clinicians diagnose a patient.

There are examples of systems designed to help automate statistical analysis, through the use of expert systems. This area has not had any major developments in the past years and the two most recent papers that cover this topic are [28] and [2]. The former describes the range of tasks that such a system needed to cover, whilst the latter describes an implementation of such a system. The emphasis of this thesis is on investigating methods to automate this analysis when the end user is the clinician however there are some features covered in [28] that are relevant to this thesis. These include the need for any such system to be able to explain itself, cater for user error, recommend the most powerful technique, adapt for data quality issues, incorporate new techniques and self document.

The market for statistical analysis tools includes specialist tools for the clinician and the statistician. However these offer little guidance on the overall model selection process. Some will recommend the best analysis based on the distributional assumptions of the data in isolation, whilst others will flag a break in the assumptions within the results outputs if it occurs. A survey of over 30 years of research on Intelligent Discovery Assistants (IDA) is provided by Serban *et al.* [71]. The article considers all the data analysis systems available and points out that they lack any kind of guidance as to which techniques can and should be used. Although the scope of the systems considered in their article is primarily data mining ones there are still some relevant insights applicable to statistical model selection, as statistical model selection can be seen as one of the tasks performed by an IDA. The authors state that the problem of model recommendation is challenging to formulate through precise guidelines. The review paper [71] aims to identify the different types of support an IDA should offer and the type of background knowledge that IDA needs to rely on. It then leverages these to assess how existing offerings address these and makes recommendations for a desirable set of requirements for future IDAs.

The processes that Serban *et al.* [71] suggest are relevant to IDA are the support for choosing the adequate algorithm and parameters at each step of the analysis process, as well as offering a template for the overall process. The provision of a graphical user interface, templates for common workflow and explanations regarding any decision or recommendations in order to provide rationale for the recommendation and support in the interpretation of the results. The ability for the system to cater for the user's prior experience by offering different levels of support is also suggested as being desirable.

Equally important to an IDA is access to the prior knowledge on the possible techniques to be applied. In [71] the authors argue that an important aspect of an IDA is the choice of the scope of the knowledge stored and how it is represented. This knowledge should include a registry of operators (statistical models would be classed as operators if the author's definition from [71] is applied), meta data (data about the data) on the input data for the analysis and properties of the operators. The authors separate the properties of the operators into external and internal ones. The external properties include the inputs, outputs, preconditions and effects (IOPE). The internal properties are related to the model type and parameters. It is the information held in IOPE that is crucial in ascertaining if and when an operator (model) is applicable, for example some models are only applicable to a target variable of type 'time to event'. The authors suggest this knowledge could be stored in hardcoded form, in an ontology or suggest the use of semantic web rule language. Expert rules about the process and case based reasoning (what analysis was done in past similar scenarios) should all be part of the prior knowledge of such a system.

The ideal properties of an IDA suggested by Serban *et al.* [71] should include: knowledge about all the available data analysis operators, automated extraction of all of the required information from the input data in order to advise which techniques are relevant, be able to rank all the possibilities based on user preferences (if there is more than one option for the specific analysis) and learning from the past. The first property was identified as a limiting factor in many of the IDAs assessed as many supported a very narrow range of techniques and were not extensible.

The ideal specification for an IDA according to [71] would be: extensible, self maintained, would include workflow execution (recommend and execute), leverage an ontology in order to store prior knowledge in a common vocabulary to avoid ambiguity, transparent knowledge that can be queried and self learning. All of these are desirable properties and guidelines that have been considered in the design of the approach to automated statistical model selection in this thesis.

A more recent paper shows some renewed interest in the automation of statistical analysis, the authors present initial work on a project they term "Automatic Statistician" [51]. The approach and type of analysis tackled is different from the one this thesis is focusing on. Lloyd *et al.* [51] focus on time series data and on analysis that is totally independent of any end user interaction, as the approach taken is to explore all the possible model options before selecting the model that best explains the data. The approach proposed in this thesis assumes that transparency and interaction with the end user will provide confidence in the model recommendations made.

A more general approach to model selection is developed by Stratton in [75] where the objective of Stratton's thesis is to explore modelling process automation in both equation based modelling (EBM) and agent based modelling (ABM) and propose a framework for automated ABM model specification. The approach proposed is also evaluated on empirical data. The notion of a model in [75] is one that does not have an obvious specification. In comparison the models referred to in this thesis are models with a statistical theoretical origin.

2.2 Argumentation

In the previous section I focused on the existing approaches to the support the automation of the recommendation of a model for a specific analysis. There are many requirements that have been suggested for such as system, and parallels have been made to the process of patient diagnosis. An additional area of research that underpins the method proposed in this thesis is argumentation as it has been proven to help in situations where there is conflicting or incomplete information, leverage a knowledge base, provide justifications for any recommendation and has been successfully implemented as part of decision support systems, specifically clinical decision support systems.

In this section I will provide an overview of argumentation in the context of artificial intelligence, and more specifically relate it to the research question this thesis aims to address. Applications of argumentation within the context of clinical decision support will also be described whilst drawing parallels to the context of statistical model selection requirements.

The field of argumentation has its roots in ancient Greek philosophy and is now of growing interest within computer science and artificial intelligence. In philosophy the objective of argumentation was to understand what was required to identify fallacies or errors in reasoning when trying to prove an argument. In artificial intelligence, argumentation can help the decision process when there are conflicting claims, incomplete or uncertain information.

In order to explain argumentation and the role it will takes in this thesis I will cover the following areas: What is an argument and how is it structured ? How is an argument generated? How can an argument be visually represented? What is an Argumentation Framework? How do you reason with a set of arguments to decide which arguments are acceptable?

The initial motivation that brought argumentation to the attention of the AI community was as a supporting approach to non monotonic reasoning. Argumentation can be beneficial when needing to reason with incomplete and uncertain information [15]. Specifically the main difference between mathematical proof and arguments lies in the definition of the argument as being defeasible. In other words even if an argument is not defeated given the information available at the time, it still has the potential of being defeated in future.

Argumentation plays a fundamental role in this thesis as it can provide structure to the interaction between the clinician, the statistician, and the available data when deciding which model to implement when analysing data.

A research question related scenario

In order to examine how the different aspects of argumentation are applicable to the research objective of this thesis I am going to use an example that is based on a typical interaction between clinician, statistician, and the data.

The situation is as follows: The clinician is looking to publish a paper on the suitability of a new operative technique, and they have access to the relevant data. The success of this new technique is measured by looking at the patient survival and comparing survival times and curves between two groups. One group would include patients on the new treatment and the second group would include patients on the conventional treatment.

In some cases the clinician may suggest a statistical approach and the statistician would ensure it is appropriate, and if this is not the case the statistician would suggest an alternative. There will be situations where the clinician does not have an analysis approach in mind. In both of these cases the choice of model would depend on a number of factors, some of which are intrinsic to the data and others that are external to it. In some cases there will be more than one appropriate statistical model to use. It is at this point that argumentation could play a key role in evaluating the reasons for and those against each different model approach, suggest the strongest one and justify this selection to the clinician.

I will refer back to this scenario throughout this chapter in order to illustrate how the use of argumentation would apply to it

2.3 Definition of an Argument

There are a range of definitions as to what constitutes an argument and a minimalistic definition is provided by Walton [84]. In this paper Walton explains that an argument can be defined in its most basic format as a set of statements or propositions and is made up of three parts: a conclusion, a set of premises and an inference from premise to conclusion. Various synonyms are used to represent these concepts. For example some additional terminology to express the definition of an argument is provided by Besnard *et al.* [17]. The authors expand the terminology to include the concept of support (equivalent to the premise) and a claim (equivalent to the conclusion or the consequent).

When relating these concepts back to the scenario introduced in Section 2.2 an argument in support of performing survival analysis could be constructed from the following elements:

1. Conclusion: Perform survival analysis.

2. Premises: The data contains survival times.

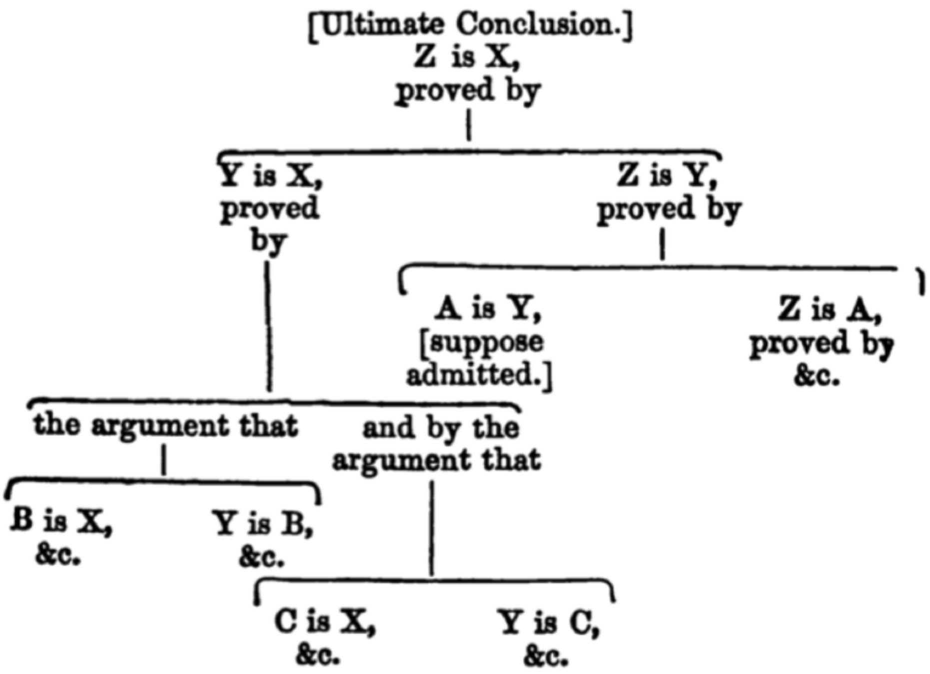


FIGURE 2.1: From Whately's Elements of Logic 1852

3. Inference from premise to conclusion: The way to analyse survival times is to perform survival analysis.

2.3.1 Argument Diagrams

The idea behind the use of diagrams as a method for argument analysis, the authors explain in [67], was first pursued by Whately in 1936. Figure 2.1 shows Whately's graph taken from [88]. His method firstly figures out the conclusion of the argument, then traces the assertions that were made in support of the conclusion. This results in a chain of arguments, represented by a diagram.

Argument diagrams are an important tool used to assist with the task of analysing and evaluating arguments by visualising the internal structure of the argument. These assist in understanding which statements are being used as premises to other statements or

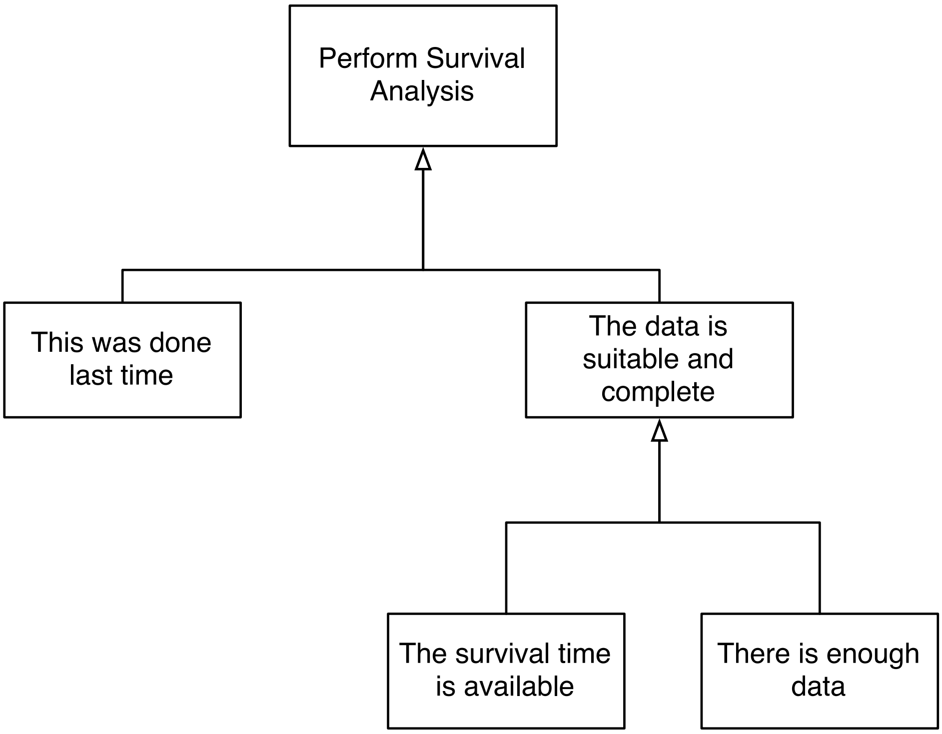


FIGURE 2.2: This is an example of an Argumentation Diagram

claims. Reed *et al.* [67] present a thorough overview of different argument diagramming methods. The authors define the argument diagrams as being made up of two main components: A set of circled numbers arranged as points; A set of lines or arrows joining these points. Each line or arrow represents an inference.

Figure 2.2 depicts an argument diagram that could arise from the scenario introduced in Section 2.2. In Figure 2.2 there is one convergent and a set of linked arguments. Linked arguments work together in support of a conclusion whilst in the case of convergent arguments each argument represents a separate reason to support the conclusion. Example 2.2 shows one diagram style, but others are possible. Following Whateley's diagram there were no major developments in the area till the 1950s and Toulmin's model.

2.3.2 The Toulmin Model

Toulmin [83] identified the importance of the layout of an argument in identifying its key components. Toulmin laid the groundwork for argument structure in his collection of essays on argumentation [83]. The Toulmin model of arguments includes these components:

- Claim: the conclusion of the argument.
- Qualifier: the strength of the argument for the claim.
- Data: the premise.
- Warrant: the inferential leap from fact to qualified claim.
- Rebuttal: circumstances that are exceptions to the warrant i.e. conditions that may defeat the conclusion.
- Backing: justification for the warrant so it is the assurance to support inferential passage.
- Qualified Claims: conclusion to be drawn if the warrant holds and the rebuttal does not hold.

The layout of Toulmin's argument components is illustrated in Figure 2.3, where the relationship between the concepts introduced above can be visualised.

Toulmin's model provides important concepts and offers expressivity in the layout and justification of arguments. It does however present some limitations. In [17] Besnard and Hunter point out that the Toulmin model is static and text based so it requires some interpretation. Its use of natural language can lead to ambiguity as different listeners may perceive different messages from the same text and as such this makes this model

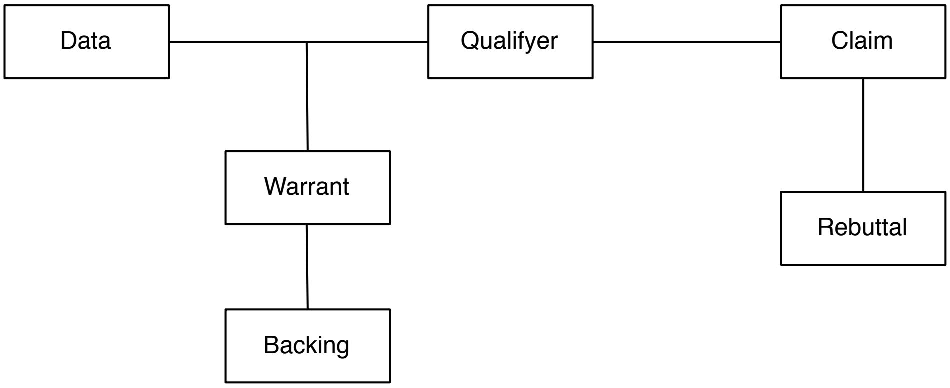


FIGURE 2.3: This is an example of Toulmin's Argumentation Layout

difficult to automate. In order to facilitate the automation a layer of ontologies or hierarchical description of concepts relevant to the domain would be required.

2.3.3 Argumentation Schemes

So far the literature review has focused on the internal structure of an argument. The internal structure of an argument may be complex and contain multiple statements in support of claims and other statements. As argument diagrams help in understanding the internal structure and dependencies of arguments, a method is needed to identify the argument type, and ultimately help in validating the claim of the argument. Argumentation Schemes [85] help in classifying different types of arguments in order to deal with each type in an appropriate manner. Argumentation schemes are forms of abstract argument that represent structures of common types of arguments, some of which are specific to the legal and scientific domain.

One of the key features of argument schemes is the list of associated critical questions (CQ). The claim that a scheme supports is presumptive and the claim is withdrawn unless the critical questions posed have been answered successfully. Below is an example of a scheme and associated critical questions.

A pertinent example of an argumentation scheme is the scheme for presumptive reasoning [11]. In this type of argument scheme given an argument we have a presumptive reason to perform an action. The presumption can be challenged or withdrawn. The following scheme is from Walton's book [85].

- Necessary condition (W1)

  G is a goal for agent.

  Doing action A is necessary for agent a to carry out goal G.

  Therefore agent a ought to do action A.

- Sufficient condition scheme (W2)

  G is a goal for agent.

  Doing action A is sufficient for agent a to carry out goal G.

  Therefore agent a ought to do action A.

The associated critical questions:

- *CQ1: Are there alternative ways of realising goal G?*
- *CQ2: Is it possible to do action A?*
- *CQ3: Does agent a have goals other then G that should be taken into account?*
- *CQ4: Are there other consequences of doing action A which should be taken into account?*

The instantiation of the appropriate argumentation scheme, in conjunction with its associated critical questions is a method of generating a set of arguments. The inference mechanism characterised by the argumentation scheme will ensure that only arguments that have not been undercut, or rebutted through the critical questions will be generated. Later in this review I will describe applications of argumentation schemes as part of decision support systems.

The process required when considering the possible model approaches within statistical model selection does not vary from situation to situation. The inference steps in the process to be taken are the same with each different research question and data. The differences lie in the models to be considered and their respective critical assumptions. The template for this inference process lends itself well to being formalised as an argumentation scheme. This is the approach that I adopted in this thesis for the purpose of generating arguments in support of the use of each model, and the argumentation scheme's critical questions will be used to rebut any of these arguments if the conditions are not met.

2.4 Reasoning with Arguments

The previous sections covered the definition, internal structure and generation of an argument, the next step is to define how arguments interact. Once arguments are constructed and validated conflicts between arguments need to be reasoned with, evaluating in light of any existing conflicts, which arguments are acceptable and finally defining the justified conclusions based on this process. In order for a claim to be accepted we need to evaluate the argument that supports it as well as any other arguments that attack the original argument of interest. In order to delve into this domain there are a few definitions necessary:

- Rebutting argument: A rebutting argument is an argument with a claim that is the negation of a conclusion (claim) of another argument.

- Undercutting argument: An undercutting argument is an argument with a claim that contradicts some of the premises of another argument.
- Defeated argument: Both rebuttals and undercuts are types of defeaters.
- Argumentation is the process by which arguments and counter-arguments are evaluated.
- Argument relations: An argument can either be supported by another argument, or it can be attacked either by another argument or by raising critical questions about it.
- Argument semantics: Represents the arguments that are acceptable.

Lets look at ways to relate the terminology introduced above in the context of the scenario described in section 2.2.

- Clinician's Argument (CA): In order to determine differences in survival curves Kaplan Meier Survival Analysis is to be preformed.
- A potential undercut: The data is so heavily censored that the results would be biased so a χ^2 analysis on 3 year survival should be performed. The premise of CA is contradicted.
- A potential rebuttal: In order determine the differences in survival curves run proportional hazards model. The claim of CA is contradicted.

A claim can be accepted if there is an argument that supports it, and the latter argument is not itself attacked by a counter argument that is not itself successfully attacked by another argument in the set. In other words a claim is accepted if its supporting arguments survive the conflict with all the other arguments in the set considered.

2.4.1 Abstract Argumentation

One of the major steps in bringing argumentation and computer science closer is the work done by P.M. Dung [31]. The main objectives of this paper were to explore ways to implement the mechanisms of argumentation in computers. The paper introduces the concept of acceptability of arguments and investigates how this theory can be used to explore the logical structure of many practical problems. Dung's focus is on abstract argumentation and as such the focus is on the relationship between arguments, not on the content of the arguments themselves. Arguments are represented as symbols.

Abstract argumentation's aim is to assign a status to each argument. Each argument can either be accepted or defeated (or refuted). The evaluation of whether an argument is acceptable or not is with respect to the given argumentation framework (or relevant collection of arguments).

One of the central concepts that Dung defines in his paper is that of an argument framework : An *Argument Framework* as a pair made up of a set Arg of arguments and $\mathcal{R}$ is a a binary relation on Arg. $\mathcal{R} \subseteq Arg \times Arg$.

$AF = \langle Arg, \mathcal{R} \rangle$

In effect an argumentation framework is a directed graph in which the arguments are represented as nodes and the attack relation is represented by the arrows. Given such a graph, one can then examine the question on which set(s) of arguments can be accepted. Figure 2.4 shows how three arguments attack each other. In this case we have three arguments A, B and C. B attacks A, but B is also attacked by C. In order to ascertain which arguments survive the conflict one will assign a status to each one as a starting point and then decide if this results in a set of accepted arguments. A formal method to define the outcome of such a process is called argumentation semantics. There are two formal approaches developed to defining argumentation semantics: labels and

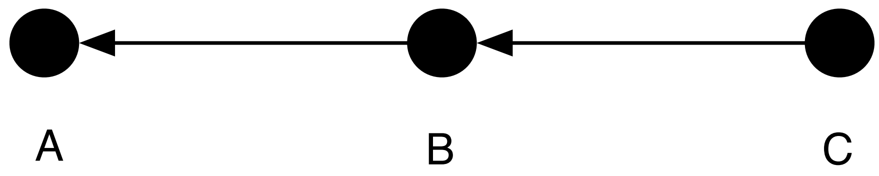


FIGURE 2.4: An example of an arguments and their relations

extensions. Detailed overviews of the different approaches with examples are in [20], [21] and Caminada also provides an introduction to argumentation semantics in [19].

Label based

In this method each argument within the set of interest is given a label. In [14] Baroni *et al.* describe the various methods used for labelling. One of their suggested sets of labels is `in`, `out` and `undecided`. If an argument does not receive any attacks then it can be labelled as `in`

- an argument is labelled as `in` if and only if all its defeaters are labelled `out`.
- an argument is labelled `out` if and only if it has at least one defeater that is labelled `in`.

Extension based

The idea behind the extension based approach is to find sub-sets of arguments within the set of arguments which can survive the conflict together. In order to do so a few concepts need to be introduced. These were all introduced by Dung in [31].

1. A set S of Arguments is said to be *conflict-free* if there are no arguments A and B in S such that A attacks B.

2. An argument $A \in Arg$ is *acceptable* with respect to a set S of arguments if and only if for each argument $B \in Arg$: if $B\mathcal{R}A$ then B is attacked by S.

3. A conflict free set of arguments S is *admissible* if and only if each argument in S is acceptable with respect to S.

4. A conflict free extension S of an argumentation framework AF is *complete* if and only if $X \in S$ if and only if S is acceptable with respect to X.

5. A grounded extension of an argumentation framework AF is a minimal complete set of AF.

6. A preferred extension of an argumentation framework AF is a maximal admissible set of AF.

7. Every argumentation framework possesses at least one preferred extension.

8. A conflict-free set of arguments S, is called a *stable extension* if and only if S attacks each argument which does not belong to S.

9. Every stable extension is a preferred extension, but not vice versa.

Dung's argumentation semantics include complete, stable, preferred and grounded semantics.

The argumentation frameworks relevant to statistical model selection will contain arguments in support of the use of different models. Each set of arguments supporting the same model will form a conflict free set. This is not informative enough, however once sets of preference arguments are incorporated and the argument framework is extended to an extended argumentation framework (EAF)[58], then the acceptable arguments according to the preferred extension semantics with respect to the EAF support the aim of the thesis. As the empty set will always be admissible yet not informative to

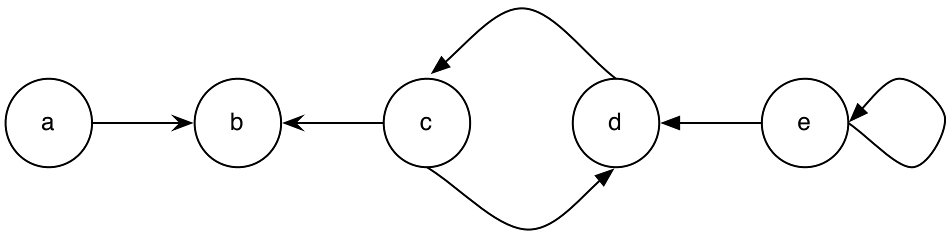


FIGURE 2.5: This is an example of an Argumentation Framework

the aim of statistical model selection, admissible extension semantics are not suited to this purpose. The proposed use of preferred extension semantics will ensure that the preferred extensions of the EAF accept as many of the arguments as they can defend, thereby including the empty set only when it is the only preferred extension. Therefore in the context of statistical model selection preferred extension semantics of the argumentation framework would be sufficient to support the aim of providing a set of suitable models.

An example

Figure 2.5 contains an example of an argumentation framework and the definitions of Dung's Abstract Framework. Given the arguments and their attack relations in Figure 2.5 then the objective is to ascertain which sets of arguments are able to survive the conflict and whether there exists only one extension or are there several solutions (sets of arguments) possible.

The conflict-free (cf) sets in Figure 2.5 are:

$cf(F) = \{\{a, c\}, \{a, d\}, \{b, d\}, \{a\}, \{b\}, \{c\}, \{d\}, \varnothing\}$

The Admissible Sets (adm) in 2.5 are:

$adm(F) = \{\{a, c\}, \{a, d\}, \{c\}, \{d\}, \varnothing\}$

$\{b, d\}, \{a\}, \{b\}$ are not admissible.

The Grounded Extension ($ground$) is:

$ground(F) = \{\{a\}\}$

The Preferred Extensions ($pref$) are:

$pref(F) = \{\{a, c\}, \{a, d\}\}$

The single stable extension ($stable$) is:

$stable(F) = \{a, d\}$

From the example:

- each AF has a unique grounded extension.
- each (finite) AF has at least one preferred extension.
- existence of stable extensions is not guaranteed.

2.4.2 The Argumentation Process

The concepts introduced so far can now be related to the process of argumentation. The process of argumentation includes the instantiation of arguments and their attack relations followed by the computation of extensions. The use of argument schemes is a method for instantiating arguments in an argumentation framework prior to computing the extensions. The knowledge required in support of the argumentation schemes (when such knowledge is required) can be held within a knowledge base. Caminada *et al.* provide a graphical representation of this process [21] that can be seen in Figure 2.6. In the first step in Figure 2.6 a knowledge base would be used to generate a set of arguments and determine their relation (attack). This would create the argumentation framework and the next step would be to determine which arguments can be accepted. Finally, based on a pre-defined criteria, the set of accepted conclusions can be identified.

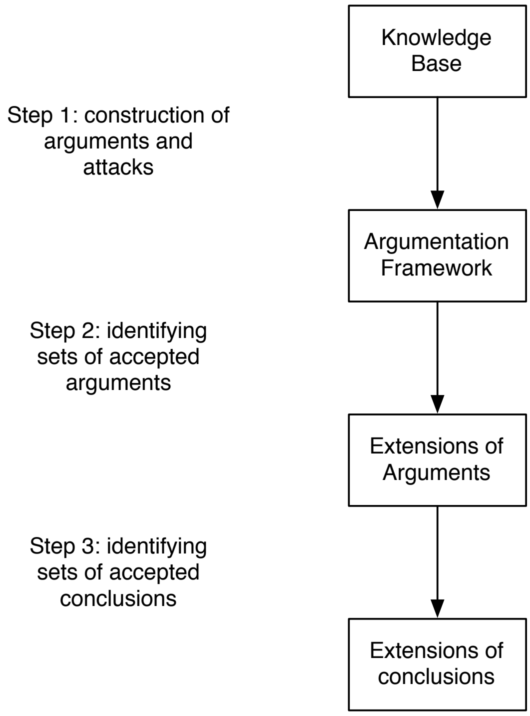


FIGURE 2.6: This is an example of an Argumentation Process

Additional approaches to the argumentation process are proposed by Atkinson *et al* [10] where the authors look at combining abstract argumentation and argumentation schemes by leveraging the desirable features of each method. In their paper [21] the authors also address the end to end process of argumentation for inference. They focus on potential issues emerging when instantiated arguments and abstract argumentation interact.

2.4.3 ASPIC

The Argumentation Services Platform with Integrated Components (ASPIC) [60] project was a joint effort aimed at achieving a consensus for theoretical models of argumentation and provides practical argumentation services based on this agreed format. An additional highly desirable output of the ASPIC project was the argumentation interchange format (AIF). The software components developed as part of the ASPIC project implements capabilities in: inference, decision making, dialogue and learning. One of the key objectives of this project was to produce a robust and portable product.

Inference within ASPIC consists of four steps according to [37]:

1. "Argument construction - tree structured arguments can be constructed from a knowledge base K of facts and a set S of strict rules of the form $\alpha_1, ...\alpha_n \rightarrow \beta$. and a set R of defeasible rules of the form $\alpha_1, ...\alpha_n \Rightarrow \beta$.

2. Argument valuation - assigns weights to arguments using validation schemes that depend on the application domain.

3. Argument interaction - is based on the binary conflict relation of attack and defeat between constructed arguments.

4. Argument status evaluation determines winning or justified arguments based on the graph of interacting arguments and using Dung's calculus of opposition [31]."

The argument source is in the form of a knowledge base of facts and rules, each having a numeric degree of belief in the range of (0,1]. A strict rule has a degree of belief of 1. A degree of belief of less than one indicates a defeasible rule. Each argument has a claim and a numeric support (0,1] that the argumentation engine uses to resolve mutual attacks between arguments. Attack and defeat relations need to be defined between arguments in order to define acceptability. Within ASPIC there are three types of attack relations: A *rebut* is a premise attack, An *undercut* is an attack on rule application and *Restrictive rebutting* ensures that an argument whose top rule is defeasible cannot rebut and argument whose top rule is strict.

The ASPIC project aimed to support an extended range of decision criteria, and decision candidates can be specified using elements of the underlying logical language. The inference components construct acceptable arguments for and against decision options, and the decision-making component processes these arguments.

> "In argumentation theory the intention is to determine if a particular proposition follows from certain assumptions even if some of these assumptions are disproved by the other assumptions. Furthermore, arguments can have relative strengths, which provide a very human-like approach to reasoning" [41]

The ASPIC framework aims to confirm that one can give a general structured account of argumentation that is intermediate in its level of abstraction, providing guidance on the structure of arguments, the nature of attacks, and the use of preferences while at the same time accommodating a broad range of instantiating logics and allowing for the study of conditions under which the various desirable properties are satisfied by these instantiations. ASPIC+ [60] is a more generalised version of ASPIC [65] that can accommodate a broader range of instantiations including flat Assumption Based Argumentation (ABA) [32] and Argument Schemes [85]. As our proposed approach makes use of argumentation schemes then the ability to accommodate for them is of relevance. ASPIC has been used as basis for several implementations of argumentation within clinical decision support and I will cover a few examples of these.

2.5 Argumentation summary

In Sections 2.2 and 2.3 I summarised the relevant research relating to the structure of an argument and showed examples of its internal structure and how it can be graphed.

I also covered the argumentation scheme as an inference mechanism to generate arguments. The process of statistical model selection would require a mechanism to generate the arguments in favour of the use of a model (or against its use) and this would lend itself well to being formalised as an argumentation scheme. There will then be a need to reason with all the arguments generated as an argumentation framework.

In Section 2.4 I focus on the process of reasoning with a set of arguments. In Section 2.6 the implementations of argumentation for decision support will illustrate how these concepts are applied.

2.6 Argumentation in Clinical Decision Support

The aim of this section is to investigate the different implementations of argumentation in the field of clinical decision support that bear relevance to my research question. The role of argumentation within this thesis resides in the decision making process that creates the recommended analysis plan based on the clinician's analysis objective, expert knowledge and the available data.

In section 2.2 I reviewed the use of arguments and argumentation as a method of formalising the process of reasoning under uncertainty in conjunction with the ability to weigh the pro's and con's of a pending decision. As a result of this argumentation is a prominent method to support decision making in clinical situations. There are a number of implementations where argumentation has been used as part of clinical decision support and these cover a wide range of clinical decisions.

An overview of the benefits of argumentation for clinical decision support is provided by Fox *at al.* [36] where they emphasise the need for a decision support system to leverage qualitative as well as quantitative inputs. Further in their paper Glasspool *et* *al.* [37] introduce their work in developing argumentation based services for biomedical applications. The authors explain that often clinicians revert to making decisions based on qualitative reasoning, as it is often difficult to acquire the numbers required to model decisions, argumentation, as a way of formalising human cognition can bring benefit to this situation. The authors initially explain and review existing applications, focusing on their limitations to explain the reasoning behind the authors' new proposed formal foundation for argumentation systems. The authors also document the lessons learnt from these implementations and explain that the safety implications in the clinical field make the development of a formal foundation a very important step. The authors' development has been based on a formal logic of argument, (LA)[48]. Contrary to standard logic in LA different arguments can simultaneously support or oppose propositions, this scenario is common in the medical domain where knowledge can be inconsistent.

The authors describe LA as including: the claim being made, the internally consistent grounds of the argument, a representation of confidence.

The authors describe a selection argumentation based decision support systems that were based on the PROforma [38] specification language. The PROforma argumentation based decision model supports natural form of explanations for and against the possible decisions. The authors describe three applications of the LA model, the details regarding the two implementations that bear relevance to this thesis are elaborated below.

- RAG (Risk Assessment in Genetics) is an implementation of argumentation for clinical decision support, this was specifically designed to help GPs with initial counselling for patients with concerns regarding specific genetic conditions. This system described in [23] offers an interface for the collection and analysis of a patient's risk based on their characteristics and their family history. The system determines a patient's resulting risk by adding the pros and the cons, these are a collection of arguments pertinent to each genetic condition that are stored in the engine of the system. Overall RAG was seen to improve clinical decisions and offer clear explanations in support of recommendations, with out any accuracy loss.

- REACT (Risk, Event, Actions and their Consequences over Time) is another example of the use of argumentation as part of a medical care planning support system. REACT is described in [39]. Decisions are rarely made in isolation and each decision or treatment will have implications on the future treatment plan and prognosis. REACT enables the patient and the clinician to map out the treatment plan and simulate the implications of different interventions. All the information is processed through REACT as logical arguments that include verbal arguments for and against and that allow for qualitative and quantitative inputs. The arguments are compiled using a range of methods including Bayes and domain knowledge. This system was trialled at Guy's Hospital in genetic cancer counselling. One of the advantages of the approach taken in this system is that it helped structure the consultation with the patient and the visualisation was also seen as especially beneficial in that situation. However it was noted that some patients may not benefit from the graphical representation, so its use as a patient facing tool would need to be considered on a patient by patient basis. It was also noted that the clinicians would need to acquire new skills to work with a system of this type.

These applications demonstrate according to [37] that only part of the behaviour can be captured, as the PROforma model is bound to a decision criteria based on aggregation. It is not possible within this model to define argument interactions, in other words arguments cannot attack other arguments. The authors develop the ASPIC model in order to address this and additional concerns raised with the PROforma based models.

2.6.1 EIRA

EIRA (Explaining, Inferencing, and Reasoning about Anomalies) is an example of an ASPIC based clinical decision support system and is described in [41]. Its subsequent enhancements are covered in [42]. The problem that EIRA focuses on is the ability to detect a patients' anomalous reaction to a medication within the Intensive Care Unit (ICU). The system's objective is to detect an anomalous reaction and to describe to the clinician the reasons why the observed patient reaction has been flagged as anomalous. The authors use medical knowledge as a basis for reasoning templates that form the basis for the argument generation. A scheme is expressed in terms of concepts derived from medical ontology. At run-time the scheme instantiates an argument by using domain information in combination with the patient's specific data. After the generation of the arguments and their interactions a graph is created illustrating which arguments are attacked or defended. This forms the basis for inferring whether the patient's response to a specific medication was anomalous.

In order to represent the exchange of arguments in the ICU domain, the authors held sessions with clinicians and captured the possible explanations as to why a particular reaction to a medication had occurred, and whether given all that was known about the patient and the medications this was an anomalous reaction. The authors then formalised the main arguments raised using the ASPIC engine. These formed the basis for the knowledge base.

In order to use ASPIC the authors completed the following steps:

1. Defined a knowledge base: this knowledge base included a set of facts and rules. Confidence in facts is expressed as support, the greater the numeric support the more confidence in the validity of the rule.

2. Defined argument interactions in line with the attack relations defined in ASPIC for rebutting, restricted rebutting and undercutting. Note that strict arguments cannot be rebutted. Under restricted rebutting an argument whose top rule is strict cannot be rebutted by an argument whose top rule is defeasible. If an argument A undercuts B, then A claims that some rule in B is not applicable.

3. Evaluated argument status: This is done using Dung's calculus of opposition [31].

In practice this system initially looks at patient data to identify what drug was administered. Then it retrieves the anticipated effects from the domain ontology. The patient data is then queried again to determine what occurred after the drug was administered, the expected and the actual effect are compared in order to determine whether an anomaly occurred.

Further to this initial implementation of the EIRA system Grando *et al.* describe in a subsequent paper [42] some further refinements to the system. This was motivated by the need to describe to the clinician the rationale behind the decision to classify a specific patient reaction as anomalous. In [42] Grando *et al.* propose to construct justifications using ontology based argumentation reasoning. This new version of the system is re-named arguEIRA.

In arguEIRA Dung's calculus of opposition [31] was used to identify and express argument interchanges made by clinicians. The obtained arguments and attack relations were modelled in ASPIC. The use of schemes allows the methods used to construct the argumentation framework to be separated from the form and content of arguments. The original anomaly detection was replaced with an argumentation based hypothesis generator module based on the schemes identified. The hypothesis generator informs a list of possible hypotheses explaining a patient's anomalous response, ordered by strength of evidence. A natural language system was planned to be used for the justifications.

The authors share some sample cases in arguEIRA including the presentation of information back to the clinician. These are available in both textual and graphical display. The feedback the authors received on this enhanced version of the system from clinicians was positive, although some clinicians did comment that the feedback was too wordy and that perhaps more graphical output would help.

The authors further discuss the importance of the separation between domain knowledge and process knowledge. The authors claim that by replacing the domain ontologies with ones from another application domain arguEIRA is re-usable. They plan to apply arguEIRA to different domains as part of their future work.

There are some parallels in the research question that underlies EIRA and the problem this thesis is addressing. The first one being the importance of feedback to the clinician, this in order for the system to be used and trusted. Another similarity is in the need for an ontology and a knowledge base in order to create the arguments and communicate the recommendation back to the clinician.

2.6.2 CARREL+

A further example of a problem that has benefited from the use of argumentation is the organ transplant system. There is a known shortage of viable organs therefore the allocation process needs to be as efficient as possible. In their papers [81, 82] Tolchinsky *et al.* describe how they extended an existing system CARREL to allow more deliberation regarding an organ's viability for transplant. CARREL is a decision support system and as such its aim is to inform and facilitate any decision, ultimately the decision will be made by a clinician. Within CARREL the decision whether an organ could be offered was based solely on the assessment of experts at the donor site. In the enhanced CARREL+ system this becomes a joint deliberation between donor and recipient agents through an argument based dialogue. The aim of this improvement was to increase availability by letting the recipient site successfully argue that the organ is actually viable.

CARREL+ focused on two main factors within the organ allocation process: Differing opinions on organ viability across different doctors and differences in receiver characteristics and circumstances. These were not part of the original CARREL system. The main extension of CARREL+ was the introduction of a Mediator Agent (MA) for managing the deliberation over the viability of an organ between the Donor Agent (DA) and the Recipient Agent (RA).

The authors used the ProClaim argument based model to implement the MA. The three main tasks for the MA were: to inform the proponent agents of the dialectical possible moves at each stage of the deliberation; to ensure submitted arguments are relevant, for example don't breach any guidelines; and to evaluate the submitted arguments to identify the winning ones and determine whether the proposed decision is valid. In order to achieve this the MA accessed four knowledge sources:

1. Argument scheme repository - this includes the argumentation schemes and with their associated critical questions. Agents construct arguments instantiating schemes and critical questions that effectively encode the full argument space which covers all possible lines of reasoning that the agent should pursue for a given issue. So the role of the argument scheme repository is not only to encode the full argumentation space but also to guide agents in exploring the full range of possible dialectical moves at each stage of the dialogue.

2. Guideline knowledge - This is to validate whether the arguments submitted comply with the established knowledge. This encodes all the medical knowledge relevant to assessing organ viability.

3. Case based reasoning engine - submitted arguments are assigned strengths based on evidence gathered from past deliberations.

4. Argument source manager - knowledge related to the agents' roles or reputation that can affect the strength of their arguments. Some transplant centres may be entitled to a greater deviation from the established criteria.

The MA submits arguments in favour of the offered organs viability using a viability scheme (VS) then the DA_i and RA_j can submit further arguments. If these are accepted by the MA, by validating against guidelines these will be arguments that attack or reinstate the argument for the particular organ's viability. The argument evaluation is done when the MA has accepted that all the arguments are valid and organises them into a graph of interacting or attacking arguments. The justified or winning arguments were then determined using Dung's seminal calculus of opposition [31].

Another interesting aspect of this system is the use of critical questions. These are not only used to instantiate the schemes, but also to challenge the agents to explore relevant areas. In the event of *"tie breakers"* MA references guideline knowledge sources to assign strengths to arguments, through a preference relation. CARREL+ was implemented on a small set of examples and it was delivered on ASPIC.

As with EIRA there are some lessons learnt that are relevant to this thesis. The presence of multiple agents bears some similarity to the situation within my research question where the deliberation on what statistical model is recommended involves clinicians, statisticians and data. Furthermore the approach used in CARREL+ ensures that the agents are encouraged to explore all options. The parallel implication of the latter within statistical model selection is to ensure that additional analysis approaches are explored if relevant.

CARREL+ is also an example of an ASPIC based implementation and in [42] Grando *et al.* discuss the differences between these two systems. Whilst arguEIRA is an automatic hypothesis generator, CARREL has been developed as an argumentation based tool for supervising and validating deliberation carried out by clinicians. arguEIRA automatically and exhaustively explores the patient's medical record and the repositories of argument schemes in order to generate all possible hypotheses and justifications. On the other hand CARREL expects the human to generate arguments for and against. A mediator agent detects the deliberations by means of critical questions associated with the exchanged argumentation schemes and then evaluates the arguments submitted by the users to conclude whether a proposed decision is valid. Furthermore CARREL does not feed back to the user in natural language.

2.6.3 DRAMA

The DRAMA agent (Deliberative Reasoning with Arguments about Actions) is the system that is proposed by Atkinson *et al.* [12, 13] to decide the best course of action for a specific patient. The approach that the authors use includes the following steps:

1. An argument scheme is used as a presumptive justification for a course of action.

> AS1: In the circumstances R
>
> We should perform action A
>
> Whose effects will result in state of affairs S
>
> Which will realise goal G

Which will promote some value V

2. This presumptive justifications must be subject to critique, and this is done through the critical questions associated with the argumentation scheme.

   Within this example the critical questions are:

   - CQ1: Are there alternative ways of realising the same effects?
   - CQ2: Are there alternative ways of realising the same goal?
   - CQ3: Are the assumptions on which the argument is based true?
   - CQ4: Does preforming the actions have a side effect that demotes some other value?
   - CQ5: Will the action have the effects described?

   The different knowledge bases used within DRAMA were implemented using simple Prolog rules. The critical questions can lead to some justifications being defeated or further actions to be considered. So for each critical question whose preconditions are satisfied, one or more arguments attacking the original justifications can be produced. These arguments may in turn be subject to the same process of critical questioning to generate counter-arguments.

3. All the arguments and counter-arguments and their attack relation are organised in an argumentation framework and Dung's calculus of opposition [31] is used to determine their acceptability.

In the specific scenario covered in [13] there were also value considerations that needed to be taken into account. The authors used an extension of Dung's framework - value based argumentation framework [16]. The implemented example also made use of six separate knowledge bases to accomplish its objective and recommend the medical treatment for the patient.

In DRAMA, similarly to the research question of this thesis, it is conceivable that more than one analysis method will be suitable, one option may be to suggest all suitable analysis. Alternatively, as was done in DRAMA, it may be an option to explore the use of value based argumentation frameworks to somehow encode the benefit of each model choice relative to each other. Another option may be to apply preferences. These will be discussed in Chapter 6 of this thesis.

2.6.4 Decision Support using Argumentation Schemes

An additional implementation of decision support using argumentation, specifically argumentation schemes and critical questions is described by Lindgren[50] where it is applied in support of clinical diagnosis. The paper focuses on the modelling of knowledge in clinical guidelines as schemes in an argumentation framework with the aim of integration into a decision support system. Preferences are also introduced in order to enrich the diagnostic process. The system provides the physician with an overview of the evidence in a patient case interpreted within different guideline contexts, represented as a set of argument schemes.

Lindgren [50] applies the use of argumentation schemes to the support of diagnostic reasoning for clinical diagnosis in the dementia domains, and in her paper she discusses the value of an argumentation based approach compared to a list of constraints approach.

> "Studies indicate that a flexible support through out the reasoning process providing motives and guidance for the situation at hand, is more useful that optimising a rulebase that only provides with suggestions of diagnostic solutions in cases where it is possible to provide one, based on the literature and available patient data " [50]

2.6.5 Reasoning with Clinical knowledge

Another challenge that emerges in the context of clinical decision support is the ability to leverage all of the existing pertinent information. This can be found not only in clinical guidelines but also in the numerous journal publications and clinical trials. In their paper [40] Gorogiannis *et al.* develop a method of representing and reasoning with biomedical knowledge. They propose the use of logical language to represent arguments and counter arguments for the relative merits of differing treatments for a specific condition. The method described supports the drawing of inferences from sets of rules that are incomplete or inconsistent. For example when the patient is not a complete match to the trial segment or different trials conclude inconsistently on what treatment is best. The authors start by defining a simple language, then based on this they develop the logic or ontological reasoning and finally define an argumentation system.

In their follow up paper [45] Hunter *et al.* further expand this system. Whilst in their previous paper the authors focused on one outcome measure for treatment comparison, in this later paper they expand the system to include multiple outcome measures. Their system (based on a logical language) uses a table of evidence as its input. The argumentation process produces a superiority graph. Their proposed process takes into account the structure of the individual argument and the dialectical structure of sets of arguments. In their set up they focus on the pairwise comparison between treatments, as even in cases where multi way comparisons are available these can always be represented in pairs.

Arguments are extracted from the evidence pertinent to the specific condition. Their framework uses Dung [31] to evaluate the arguments once they have all been generated from the evidence. For example for two treatments $t_1, \ldots, t_2$ the following claims could be the case:

- $t_1 > t_2$ evidence supports the claim that treatment 1 is superior to treatment 2
- $t_1 \sim t_2$ evidence supports the claim that treatment 1 is equivalent to treatment 2
- $t_1 < t_2$ evidence supports the claim that treatment 1 is inferior to treatment 2

The authors further introduce the concept of preferences in the process to take into account the relative benefits of the treatments begin considered. This is done though the use of preference argument frameworks (PAF) as proposed by Amgoud *et al.* [4], and defining preference relations over sets of benefits (from the different treatment options).

The relevance on the latter two papers to this thesis is in the translation of the statistical knowledge on model selection into argumentation. Furthermore the use of preferences to represent the relative benefit of different model options may be of value to the research question, and will be discussed in Chapter 6 of this thesis.

2.6.6 Argumentation for Decision Support - Summary

The common themes in the methods described in Section 2.6 are those of benefits to the end users (typically the clinician), with no compromise on accuracy. However in collating the drawbacks that have been observed with these systems the majority relate to the presentation of the information back to the user (clinician or patient) either directly or facilitated by a consultation with a clinician, highlighting the need for specialist training of the clinicians and that of tailoring the output in line with clinician feedback.

In the context of this thesis these implementations show examples of successful argumentation based decision support systems and confirm that these can be accepted for use by clinicians. However none of the implementations focus on the use of argumentation for the purpose of automating statistical analysis, the area this thesis and its contributions is focused on. The lessons learnt from these implementations are relevant to this thesis. There are multiple aspects addressed within these sample implementations that will help shape the proposed system and contributions of this thesis.

2.7 Preferences

Preferences are an important element in decision making, especially collective decision making, notably in situations where preferences cannot be expressed in a binary way, distinguishing good alternatives from bad, or easily enumerated in a list; then dealing with them becomes a non trivial issue.In the context of statistical model selection preferences offer a method of expressing the relative strength of support for the use of one model compared to the use of another one. In their review article Domshlak *et al.* [30] discuss different approaches to representation, processing and learning of preferences within the AI domain. The authors explain that the origins of the representation of preferences was in the field of economics, specifically decision theory and as such much emphasis was placed on the utility or value that underlies the different decisions under differing conditions.

Eliciting preferences directly in terms of a value function, even if there is an underlying ordinal scale is difficult in the context of AI. The authors further discuss the option of providing information about preferences in separate pieces, such as binary preference relations. The downside of this approach is that this does not scale well as the number of pairwise comparisons grows. Domshlak *et al.*. [30] describe the use of preference statements as an alternative approach. These can describe preferences in a local and contextualised manner. These could be in graphs through logical representation. The research has focused on compact representation of preferences, mainly as graphical representations. The authors draw some parallels between the preferences communicated as pieces of information and knowledge bases. They argue that they display the same problems and concerns with regards to reasoning, revising and fusing different points of view. These three aspects are important to the use of preferences in statistical model selection within the methodology proposed.

Domshlak *et al.* [30] provide an overview of the different methods for preference representation and preferences. The options covered for preference representation include graphical representations such as CP-nets and logic based representations. The latter can be represented in different propositional logic languages that will vary according to the nature of the pre-orders that can be encoded and the compactness of the expressed preference relation.

In their article Coste-Marquis *et al.* [22] assess the expressiveness and succinctness of different preference representation languages. The authors conclude that there are differing levels of succinctness but there is no language that is always more succinct. Other factors should determine the choice of language in which to represent the preferences. In our case a good starting point are the languages that are used by existing implementations of argumentation that leverage preferences, none of those make use of cp-nets.

In the context of argumentation the use of preferences in decision support offers a way of representing the strength or priority of an argument. Preferences can also be used to quantify the quality or uncertainty underlying an argument. In their article Amgoud *et al.* [4] introduce the concept of a preference based argumentation framework(PAF) as a way to leverage preference relations into an argumentation framework , as a refinement of Dung's acceptability calculus [31]. The authors note that a weakness in Dung's definition of acceptability is it disregards the quality of the argument. Their proposal for preference based argumentation combines the preference relations between arguments with the defeasibility relations, in other words preferences between arguments are considered at the same time as attacks between argument to determine if the attack is valid.

Amgoud *et al.* [3] further develop their proposed approach by introducing preference based acceptability. Their proposal is that arguments are acceptable if they are not attacked or if they are attacked they are preferred to their attacker and therefore defeat the attack. The extensions will contain un-attacked arguments and arguments which are preferred to their attackers. The latter are arguments that defend themselves against their defeaters.

In [6] Amgoud *et al.* reviewed the methods that incorporated preferences and argumentation ([3], [58] and [16]) and concluded that these could potentially lead to unintended results. The potential inconsistency is illustrated through the use of a simple example where it is clear that one argument is not as strong as the other despite defeating it. Their proposed new approach takes three elements: a set of arguments A, an attack relation R and a total or partial preorder. Two requirements need to be satisfied by any extension, the first requirement ensures that the extensions returned by the new framework are conflict free, the second one captures the idea that that an attack fails in case the attacker is weaker than its target. As a result of this they proposed a new preference based argumentation framework (PAF) able to deal with these new requirements.

Amgoud *et al.* further refine PAF [7] by distinguishing between two separate roles for preferences; repairing the attack relation between arguments and refining the evaluation of arguments. Their new proposed abstract and general framework handles both of the preference roles. A critical attack is defined as one which emanates from an argument which is less preferred that the argument is attacks. Their proposed method to handle the dual role of preferences simultaneously starts by inverting the arrows of the critical attacks, then computes the extensions, and lastly applies a refinement on the set of extensions in order to select the best one. The authors do continue to make an assumption that the preferences are a full or partial pre-order.

An additional approach to handling preferences that may hold some contradictions is offered in Amgoud *et al.* [5], where the authors propose an extension contextual preferences argumentation framework (CPAF) to an early implementation of PAF to enable it to handle multiple points of view on an inconsistent knowledge base. Amgoud *et al.* define the concept of contextual preferences as preference orders that depend on a particular context. For example if there are two arguments each in support of the use of a different model m_1 and m_2. If one model is preferred to the other in the presence of one context, but the opposite is true with regards to a different applicable context then this is the situation covered in the paper. The authors propose that the order of importance of the contexts will determine which is the stronger preference.

An approach that is closer to the way preferences are interpreted in economics is presented in Bench Capon [16] where he proposes value based argumentation (VAF). In VAF a Dung framework is augmented with values and orderings so that an attack of m_i on m_j is successful only if the value obtained by running m_i is greater than the one promoted by running m_j.

A different approach to handling preferences in the argumentation process is provided by Modgil [58], in his article the author proposes Extended Argumentation Frameworks (EAF), a novel approach involving the extension of Dung's argumentation frameworks to include arguments on the preferences between arguments (meta-level arguments), (EAFs).

The assumption made on preferences is that they are pre-specified, but may be contradictory as they can vary according to the situation and in time therefore there is a need to argue about as well as with preference information. Preferences are also treated as arguments, but not as arguments that attack other arguments but rather as arguments that attack the preference of one argument over another one through a second attack relation. The methodology proposed in [58] preserves the abstract nature of Dung's approach [31] whilst accommodating meta-level arguments expressing preference between arguments.

In their article Modgil *et al.* [59] further generalise their ASPIC+ framework by adopting a new method for evaluating extensions. This proposal differentiates between attacks and defeats, the former are used to define conflict free sets whilst the latter as obtained by applying the preferences determines the acceptability of arguments. The authors state that they plan to extend this to structure EAFs.

The main difference in how preferences are leveraged between PAF and EAF is as follows: preference based argumentation (PAF) [6] incorporates the preferences simultaneously to evaluating the argumentation framework, whereas the extended argumentation frameworks (EAF) [58] adds a meta-layer of arguments which encapsulates the preferences between arguments.

In [27] Cyras presents a simple example of common sense reasoning involving preferences and implements it in a variety of formalisms, including ASPIC+, Assumption Based Argumentation with preferences and abstract argumentation. The problem is complemented by a survey of human responders on the decisions made given the same example. The author compared the 'rational' choice and the 'intuitive' solution that the human respondents opted for most often. The resulting choice when applying the various argumentation formalisms to the problem produced the 'rational' or other choice, but not exclusively the 'intuitive' one. The author calls for further discussion on how deal with preferences in such situations.

An example of an implementation of argumentation with preferences is described by Croitoru *et al.* [26]. The EcoBioCap project's aim is to support the selection of choice of food packaging for EU consumers by aggregating the preferences of multiple parties. Croitoru *et al.* introduce a logic of preferences within their argumentation framework and describe how arguments and argument related concepts are obtained. The authors propose a logic able to both express and reason with expert claims over preferences. This provides an example of an ASPIC+ system with preferences and argumentation schemes, specifically the argument scheme from expert opinion.

Another relevant paper is by Hunter *et al.* [44] where the clinical preferences are used as part of the argumentation process. In this paper the aim is to offer the clinician the facility to aggregate evidence whilst taking into consideration the clinician's own assessment of the strength or weaknesses of each item of evidence. A clinician's preference may stem from the source of the evidence, as this paper deals with medical evidence in support of different treatments the sources of evidence can include Meta Analysis and Clinical trial results. In the methodology proposed in [44] the preferences are placed on the evidence that is used to evaluate the arguments, not on the arguments themselves. The authors define a preference rule as a condition on a pair of arguments, and introduce the concept of a Expressed Preference Scheme (EPS) to capture each clinician's preference for each set of evidence. This method was evaluated by means of an actual trial with clinicians. The difference between our situation and the scenario considered in this paper is that in our case the preferences are not completely dependent on the clinician's view.

2.8 Summary

In this chapter I have presented an overview of the main topics of research relevant to this thesis. Section 2.1 was devoted to the topic of expert systems and specifically their use automating the process of analysing data. The conclusion from this part of the review revealed that the specific challenge this thesis is focusing on has not been addressed in past work, some aspects of it have but not in its entirety. The literature reviewed also researched the desirable features of a system aimed to help with the analysis process. The main conclusions related to the importance of the transparency of the process, the need for a knowledge base that is flexible and expandable and an approach that leaves room for interaction (not a black box).

Sections 2.2 and 2.4 have focused on the definition and generation of arguments, reasoning with arguments argumentation and the argumentation process. This covered the process from the notion and instantiation of an argument to reasoning with a set of arguments. Section 2.6 explored different applications of argumentation to decision support. The final aspect covered in this literature review is preferences which covered preferences in the context of decision support as well as in argumentation. In Section 2.7 different approaches for the use of preferences in argumentation are explored. The approaches relevant to the challenge of statistical model selection, as addressed in this thesis, need to be able to cater for multiple and conflicting sets of preference orders. The main approaches of direct relevance to the challenges of statistical model selection are PAF [6] , CPAF [5] and EAF [58]. Relevant implementations of argumentation based decision support systems with preferences were also covered in Section 2.7. The role of preferences will be explored in more detail and applied in the context of statistical model selection in Chapter 5.

My research and the problem tackled within this thesis touches on the aspects covered in this chapter and the key points relevant that will underpin the contributions in forthcoming chapters are:

- The use of an argumentation scheme to instantiate the arguments with associated critical questions.
- The need for a supporting knowledge base to hold all of the rules derived from statistical theory.
- The need for a method of reasoning with all of the arguments generated, each having a different conclusion as to what model to use.
- The requirement to both interact with the user as the arguments are instantiated, as well as the feedback reasoning as to the recommendations made.
- The use of preference orders within the argumentation framework.

Chapter 3

Arguments for statistical model selection

In this chapter I will describe the process of statistical model selection, introduce survival analysis, discuss how argumentation can be used in decision support to inform the decision maker as well as describing the applicability to statistical model selection. The aim of this chapter is to set the scene and the context for the latter part of this thesis that will introduce my original contributions.

Section 3.1 delves into the details of the process of selecting a statistical model when given a research question and relevant data. Section 3.2 explains how an argumentation system can be used for decision support and how the argumentation models inform the decision maker. The first contribution of this thesis is introduced in Section 3.3 which defines the required structure and contents of the knowledge base that will support the proposed model selection methodology.

3.1 The process of statistical model selection

Analysing available data in support of a research question or hypothesis necessitates some statistical approach to ascertain whether the effect observed in the data is significant or not. In a simple example one may have access to data on number of patients admitted or the number of students passing a test, and one may wish to answer the question: "Is there an increase between last year and this year?". Simply comparing the number of patients admitted each year or comparing the pass percentage of students on its own does not provide an answer to the question asked. A difference in the values, however large or small, can be attributed to random error and as such provides no confident evidence that there is indeed a difference between the years.

It is at this point that the statistical model selection problem arises. Each of these situations can be supported through the use of the most appropriate statistical model or approach. In some cases there will be more than one possible approach to statistically test the research question or hypothesis. A trained statistician is able to make these considerations and tests to result in a recommended analysis model or approach.

As highlighted within the introduction of this thesis, as there is an ever increasing amount of data available to clinicians to test hypothesis, coupled with easily usable statistical functionality being offered in standard off the shelf spreadsheet products there is a need to offer a method to guide, support and justify the selection of one statistical approach over another. This can be done by consulting a statistician, but this involves additional effort and time. An automated recommendation and justification in support of the most suitable model would ensure this does not add effort and time.

The choice of model in support of a research question will involve knowledge on what model approaches achieve the type of objective specified by the research question and the data available. Additionally the statistical theory underpinning the model approaches (I will refer to these as models) dictates conditions under which models cannot be used and conditions under which models may not perform at their best. The former are know as critical assumptions and the latter are context domains. These conditions can be tested either by querying the data or by eliciting the information from the domain expert of the data (in most cases the clinician).

3.1.1 Introducing the running example in detail

In order to illustrate the process of selecting an appropriate statistical model I will be introducing a simple example based on a freely available data set. The data set is called `ovarian` and it contains the data collected in a randomised trial comparing two treatments for ovarian cancer [33]. Table 3.1 shows the attributes available in the data for all 26 patients:

attribute name	attribute description
futime	survival or censoring time
fustat	censoring status
age	in years
resid.ds	residual disease present (1=no, 2=yes)
rx	treatment group (1,2)
ecog.ps	ECOG performance status (1 is better)

TABLE 3.1: `ovarian` data: column names and description

Given this data a clinician may be interested in testing the following hypothesis on the `ovarian` data:

Hypothesis 1

Is there a difference in patient survival between treatment 1 and treatment 2.

In order to test Hypothesis 1 then there is a need to select and apply a statistical model. The measure that determines survival is the survival time, in this data this is contained in the column `futime` and the censoring status is in the column `fustat` which determines if the event of interest has occurred. This type of analysis is *survival analysis* and the analysis objective is *survival.* The next section explains *survival analysis* and illustrates the process through the use of the `ovarian` data with the aim to test Hypothesis 1.

3.1.2 Introducing survival analysis

Clinical databases can include a wealth of data representing the length of time elapsed between specific events in a patient's disease progression. Common examples are the time elapsed between diagnosis and death or diagnosis and relapse. This time aspect of this data requires a special type of statistical analysis which is aimed at estimating the *survival function.*

There are instances where a patient is lost in follow up or has not experienced the event at the time of analysis, this is referred to as *censoring.* In the context of the case studies in this thesis a patient is censored if they are still alive, or still alive and disease free at the time of the last recorded follow up. This is referred to as *Right Censoring. Left Censoring* occurs when a patient is not followed from the onset of the disease, however this is not a feature of the case studies included in this thesis, but an aspect to consider in the future.

The aim of the analysis of survival data is to estimate the survival function.

$$S(t) = Pr(T > t) \tag{3.1}$$

where T is the survival time. An example of a survival curve (an estimate of the survival function 3.1) is in Figure 3.1. A graph of $S(t)$ is the Survival Curve.

When there are no censored observations the calculation of S(t) is simply the proportion of survival times that are greater than t. However, in cases where there are censored observations this becomes more complex. In [47] Kaplan *et al.* introduces the *Kaplan-Meier* non parametric estimator. This involves ordering the survival times from smallest to largest and applying equation 3.2.

$$\hat{S}(t) = \prod_{j|t_i \leq t} [1 - \frac{d_j}{r_j}] \tag{3.2}$$

Where r_j is the number of patients at risk just before time $t_{(j)}$ and d_j is the number that experience the event of interest at $t_{(j)}$. Censored patients at $t_{(j)}$ are included in r_j, so if a patient is lost to follow up at that time they still count as at risk until that time.

In order to assess whether there is a difference in survival times between different groups of patients it is useful to plot a survival curve. The difference in survival curves can be formally tested using the *log-rank test* [53] which compares the observed number of events occurring at each particular time point for each separate group to the number expected if the survival curve is the same in each group.

This technique is very frequently used in clinical data analysis and as such is very relevant to *Head and Neck* data. Table 3.2 contains the data from the `ovarian` data example that I will be using throughout this thesis. The patients were all followed from the diagnosis so there is no *left censoring*, but some patients were still alive the last time they were followed up so they were treated as *right censored* and their survival time was treated as "survived at least till their last follow up".

	futime	fustat	age	resid.ds	rx	ecog.ps
1	59.00	1.00	72.33	2.00	1.00	1.00
2	115.00	1.00	74.49	2.00	1.00	1.00
3	156.00	1.00	66.47	2.00	1.00	2.00
4	421.00	0.00	53.36	2.00	2.00	1.00
5	431.00	1.00	50.34	2.00	1.00	1.00
6	448.00	0.00	56.43	1.00	1.00	2.00
7	464.00	1.00	56.94	2.00	2.00	2.00
8	475.00	1.00	59.85	2.00	2.00	2.00
9	477.00	0.00	64.18	2.00	1.00	1.00
10	563.00	1.00	55.18	1.00	2.00	2.00
11	638.00	1.00	56.76	1.00	1.00	2.00
12	744.00	0.00	50.11	1.00	2.00	1.00
13	769.00	0.00	59.63	2.00	2.00	2.00
14	770.00	0.00	57.05	2.00	2.00	1.00
15	803.00	0.00	39.27	1.00	1.00	1.00
16	855.00	0.00	43.12	1.00	1.00	2.00
17	1040.00	0.00	38.89	2.00	1.00	2.00
18	1106.00	0.00	44.60	1.00	1.00	1.00
19	1129.00	0.00	53.91	1.00	2.00	1.00
20	1206.00	0.00	44.21	2.00	2.00	1.00
21	1227.00	0.00	59.59	1.00	2.00	2.00
22	268.00	1.00	74.50	2.00	1.00	2.00
23	329.00	1.00	43.14	2.00	1.00	1.00
24	353.00	1.00	63.22	1.00	2.00	2.00
25	365.00	1.00	64.42	2.00	2.00	1.00
26	377.00	0.00	58.31	1.00	2.00	1.00

TABLE 3.2: Attribute names and contents for the full `ovarian` data

Although the *Kaplan-meier* method is non-parametric there are some assumptions that can affect the validity of the result:

- external factors - there may be a confounding difference between the different groups or strata that is not represented within the data. This information could be gleaned from the clinician. The most common example of this is the lack of independence in censoring. This situation arises when there is a different censoring pattern between the two groups, for example if one treatment group of

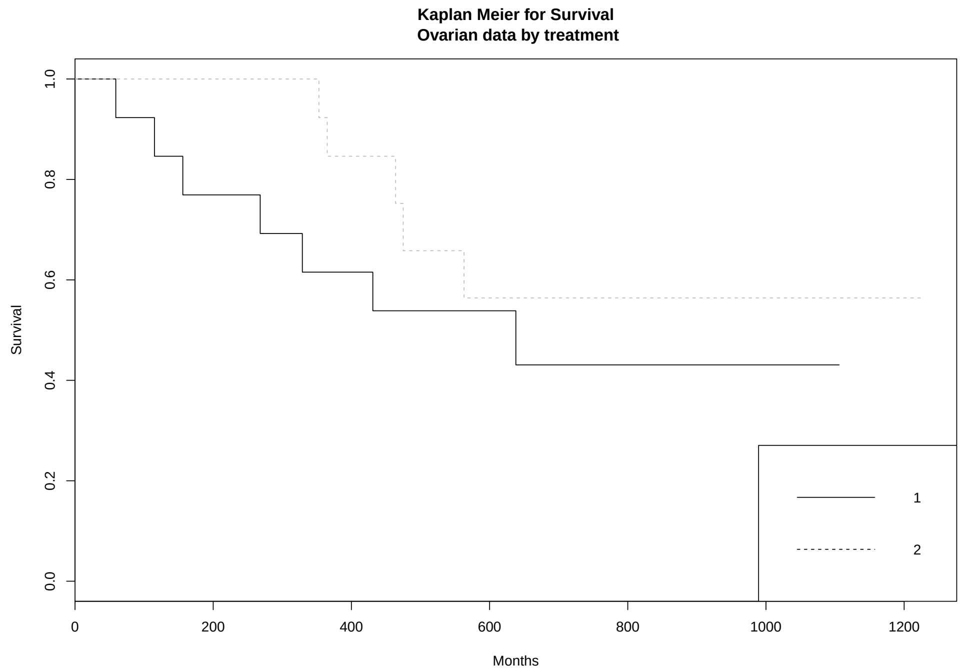


FIGURE 3.1: The Kaplan Meier Survival curves `ovarian` data

patients were to be much more likely to be transferred to another hospital, hence lost to follow up.

- heavy censoring - within the *Head and Neck* cancer domain this is a common feature of the data as the survival prospects are very good. A high percentage of censoring in the data can cause the *Kaplan Meier* survival curve estimates to be biased. The definition of heavy censoring is not rigid, but generally heavy censoring occurs if 70% or more of the patients have not experienced the event of interest. Heavy censoring does not preclude the use of this model, but potentially can cast some uncertainty on the robustness of the results.

Prior to applying the *Kaplan-meier* model to test Hypothesis 1 the assumptions listed above should be tested to ensure that they hold. The first assumption is one that requires the clinician to confirm that there are no external factors to be considered for this analysis, but in cases where there are some confounding aspects not represented in the data then this would require assessing the impact of these on the integrity of the analysis. The second assumption can also be validated with the clinician. The last assumption to be validated in order to use *Kaplan-meier* is the extent of censoring, in the `ovarian` data set there are 14 patients who are still alive at the last follow up, this results in a censoring rate of $\frac{14}{26} = 0.54$ which is considered light or mild, not heavy.

Having checked that the assumptions hold applying the *Kaplan-meier* model to the `ovarian` data to test Hypothesis 1 results in a *p-value* of 0.303 which makes the difference between the survival curves for Treatment 1 and Treatment 2 not significant at $\alpha = 0.05$.

The other common Survival Analysis method is *Cox Proportional Hazards* and it models the *Hazard function*, the method is described in detail in [24]. *Cox Proportional hazards model* is a semi-parametric method that also allows the introduction of numerical continuous covariates within the model. For example, it is through the *Cox Proportional Hazards* model that one can assess whether a patient's weight or age has any effect on the patient's survival. The *Hazard function* is defined as the probability that a patient experiences the event of interest in a small time interval s given the patient has survived up to the beginning of this interval. The *Hazard function* can be estimated as the proportion of patients experiencing the event of interest in an interval per unit time given that they have survived to the beginning of that interval. The *Hazard function* can increase, decrease or remain constant or have a more complex profile. The *Cox Proportional Hazards* model uses regression parameters as in *Generalised Linear Models*.

$$h(t) = h_0(t)exp(\beta^T x) \tag{3.3}$$

where β is a vector of regression parameters and x is a vector of covariate values.

The parameters β are estimated by maximising equation 3.4.

$$\sum_j \log \frac{exp(\beta^T x_f)}{\sum_{i \in r(f)} exp(\beta^T x_i)} \tag{3.4}$$

The parameters in a *Cox Proportional Hazards* model are interpreted in a similar fashion to those in a regression model.

There are some assumptions that must be met prior to the use of *Proportional Hazards*:

- lack of independence in censoring - This situation arises when there is a different censoring pattern between the two groups, for example if one treatment group of patients were to be much more likely to be transferred to another hospital. This is an assumption shared with *Kaplan-meier*.
- proportional hazards - The assumption of a constant hazard ratio is called the *Proportional Hazards assumption*. The *hazard functions* of any two patients are assumed to be constant multiples of each other.

As this is the same data as the one used to apply *Kaplan-meier* then there is no need to test the first assumption. The second assumption, one of proportional hazards will need to be tested. This can be achieved in two ways:

- visually - if the survival curves do not intersect (there are additional visual diagnostics)
- analytically - by testing for any significant time dependent covariates or by testing the residuals (the formulae are beyond the scope of this thesis)

The proportional hazards assumption for the `ovarian` data has been tested and it holds. This equates to there being no significant time dependent covariates. Applying the *Cox Proportional Hazards* model to testing Hypothesis 1 results in the effect of the different treatment on patient survival to be not significant (*p-value=0.305*). This confirms the findings of the *Kaplan-meier* model.

An additional approach to analysing survival data is the use of *Weibull* model [86]. This model is beneficial when certain distributional assumptions about the survival time are present. The assumptions the *Weibull* model relies on are:

- lack of independence in censoring - This situation arises when there is a different censoring pattern between the two groups. This is an assumption shared with *Kaplan-meier*.
- distribution of the log-log curve - The estimated *"log:log"* (log of the time attribute vs. log log of the survival curve) lines in the graph produced should be linear if the Weibull model is appropriate.

The second assumption does not hold for the `ovarian` data, as such the *Weibull* model is not one to consider or use in this case.

There are many additional models for survival analysis but for the purpose of this example we only chose to consider three. Of the three models considered in this example there were two models that were possible (*Kaplan-meier, proportional hazards*), and in this case both confirmed that there is no significant difference in survival times for patients in treatment 1 compared to treatment 2 (Hypothesis 1).

It is also possible to assess whether one model is more suitable that the other one by taking into account some additional conditions. A statistician may assess the situation by discussing the objective of the analysis with the clinician. If the objective of the analysis extends beyond testing Hypothesis 1 but also looks to produce a benchmark to be used to estimate survival time for future patients then the recommendation would be to use *proportional hazards* as it facilitates predictions. Another consideration relates to the different models' resilience to censoring. These are considerations that are discussed and assessed when planning the statistical analysis. There are many additional approaches to analysing survival data, however for the purpose of this thesis I am focusing on the most commonly used models.

In the example introduced the research question r is the hypothesis being tested (Hypothesis 1). The objective of such a research question is *survival analysis* (this can be represented as $O = o_s$). The knowledge required to ascertain which models are possible in this situation would need to contain information on the relations between models (m) and objectives (o). In the example the content of the knowledge base would document the link between *Kaplan-meier*, *proportional hazards* and *weibull* and the objective of analysing survival data. The knowledge base would also need to contain the relations between each model and their underlying critical assumptions. For example the assumptions of proportional hazards needing to be met in order to support the use of a *proportional hazards model*.

3.2 Proposed application of Argumentation to Statistical Model Selection

In Chapter 2 of this thesis various examples were cited for the use of argumentation in decision support. The examples were taken from clinical settings, however as mentioned in Section 2.1 there are similarities in how a clinical diagnosis is made and how a statistical model is selected. An argumentation based system can provide decision support by offering a framework to evaluate the pros and cons of a decision.

In the context of statistical model selection an argumentation system can support the structured inference from the user's research question to possible options for models to use for the analysis. Furthermore an argument can provide more than just a vote of support of a specific approach, as the knowledge and inference process used to generate an argument can also be retained to inform decisions.

A set of arguments generated supporting the use of a specific model can offer a documented reason for the support of any recommendation in favour of the use of the specific model. For example the set of arguments in support of the use of a model in a given situation could include an argument in favour of the use of the model as it achieves the desired objective and an argument in favour of the use of the same model as its critical assumptions hold. Similarly a situation can arise where the first argument in support of the use of the model as it achieves the objective is attacked by an argument against its use as its critical assumptions don't hold. The type of argument and the source of its instantiation is important in the context of statistical model selection as it forms part of the justification for any recommendation made.

The proposed approach of this thesis is based on argumentation schemes supported by a statistical knowledge base (SKB) containing all the required model information. In future chapters preferences will be incorporated into this. The argumentation schemes leverage the statistical knowledge base, the data and the clinician's knowledge in order to instantiate the arguments in support of the use of a model. This results in an argumentation framework containing arguments in support of the use of one or more models.

In most situations the desired output is a recommended model or set of models that are the most suitable one given the research question, the data and context and as such this is reflected in attack relations between each argument that supports the use of a different model. The research question and the data are problem specific, the SKB is domain specific and the argumentation schemes and their associated critical question are domain and problem independent. This is illustrated in Figure 3.2.

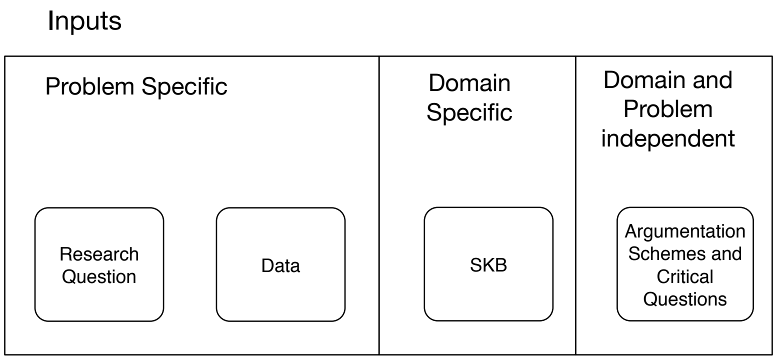


FIGURE 3.2: Overview of the statistical model selection inputs

- Statistical Knowledge Base (SKB) - This will hold all the definitions of the relationships between the research question type (R), objectives (O), models(M) and the assumptions(A).

- Argumentation Schemes - The proposed method in this thesis involves an argumentation scheme for model to consider on the grounds of achieving the objective. When this scheme is instantiated it will leverage the additional argumentation schemes as part of the critical questions in order to identify potential under-cuts, rebuttals or undermines.

- The data - The assumption is made that the data is relevant to the domain of the clinician, includes enough scope for research and is in analysis ready format.

- Preferences and context domains - This information will be elicited from the clinician and then held in a knowledge base. The contextual domains will be part of the knowledge base as they are derived principally from the known properties of each of the different models. These will be discussed in Chapter 5 where my proposed methodology will be extended to account for these.

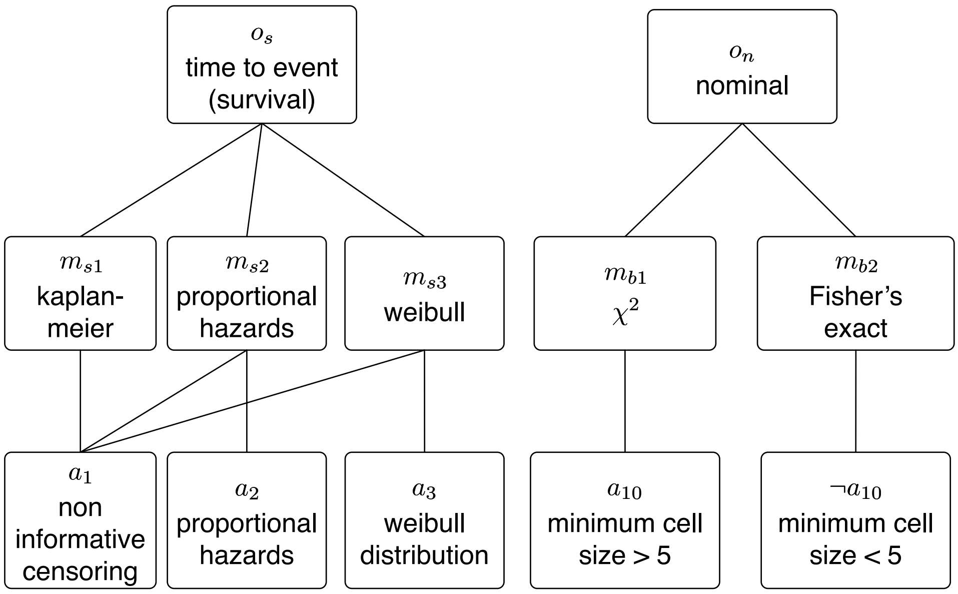


FIGURE 3.3: The Statistical Knowledge Base relevant to the `ovarian` example

The Argument Schemes are instantiated on the data and incorporate the clinician's knowledge to produce a set of arguments in support of or against the use of each of the models in scope.

3.3 The Statistical Knowledge Base

The statistical knowledge base (SKB) includes all of the relations between the research question type (R), the objectives (O) , the models (M) and the assumptions (A). The SKB will hold facts linking R, O, M, A in a way that will support the queries from the argumentation schemes and their critical questions. The SKB will hold multiple research question types and each will be linked to the objectives O that can fulfil that research question R, models M will de defined and will be linked to the respective objectives they are suitable for, and for each model the critical assumptions will be defined.

The relations and contents of the SKB are derived from statistical theory and best practice, these are the defeasible and strict rules, the defeasible premises or rules will be validated through the use of argumentation schemes. An example of the elements of the SKB can be seen in Figure 3.3, this is pertinent to the `ovarian` example. Figure 3.3 illustrates the contents of the SKB for the objectives of *time to event (or survival)* analysis and *nominal analysis (or table analysis)*. From Figure 3.3 it can be seen that there are three models suitable for an objective of type o_s and for each of these three models (m_{s1}, m_{s2}, m_{s3}) there are different critical assumptions, for example model m_{b2} relies on two critical assumptions a_1, a_2 the former (a_1) also being an assumption to both m_{s1} and m_{s3}.

The data used as a basis for answering a research question or hypothesis is assumed for the scope of this thesis to contain one row per entity of interest (e.g. patient or surgery) and multiple columns of data. The research question will involve one target variable and a set of one or more explanatory variables. The type of the target variable will dictate the type of objective of the analysis. Variables can be of type: Nominal, Ordinal or Interval. There are also some special cases such as *time to event*, as a special case of *interval* variable and *binary* as a special case of *nominal*. Figure 3.4 illustrates the different data types, their related objectives and for each objective the list of models that could be applicable. From Figure 3.4 it can be seen that *time to event* attributes are a special case of *interval* attributes and these could be analysed using models such as *Kaplan-meier, Proportional Hazards or Weibull*. Note that the methods listed in *italics* within Figure 3.4 are not included in the scope of approaches documented in this thesis, these could be added by expanding the knowledge base in the future.

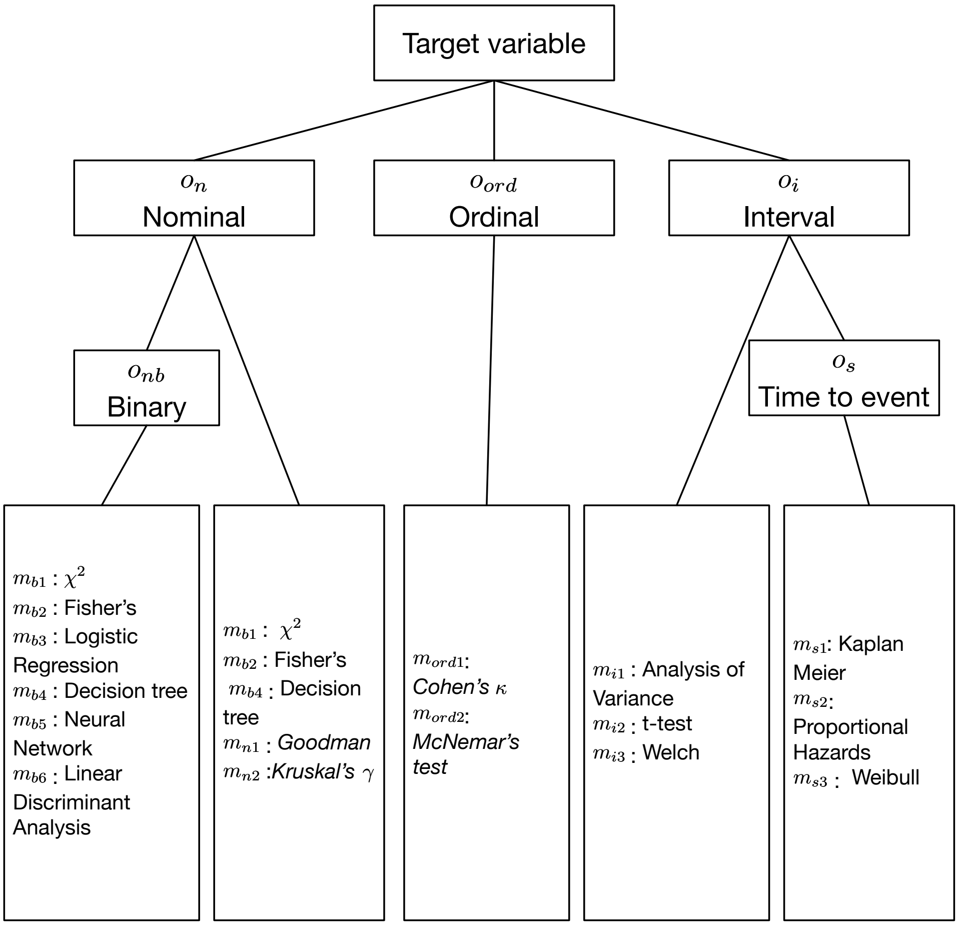


FIGURE 3.4: The different data types and their related objectives

A key benefit of the architecture proposed within this thesis is that it differentiates knowledge into domain and problem specific information to be provided by the clinician, the problem independent domain specific statistical knowledge base and problem and domain independent argumentation schemes. This facilitates maintainability of the approach and makes the methodology adaptable to different domains.

Within the process of statistical model selection there are entities such as `model` m_i, `assumption` a_p, and `objective` o_q. The `model` is the actual statistical technique, examples of this are linear regression [66] or kaplan meier survival analysis [47]. The `assumption` is a condition that needs to be tested, there are different features for each assumption and these include whether the assumption is critical or not, or whether the assumption is one that can be tested by querying the analysis data or whether the assumption is a question for the end user. The `objective` is derived from the type of the dependent variable. The example introduced in Section 3.1.1 has an objective of survival as the target variable of interest is of type *time to event*.

The main entity and the one that decisions need to be made on is the `model`. The knowledge will all relate to the model. A `model` has `assumption` (it can have more than one) and can fulfil a set of `objectives`. Initially the non critical assumptions will not be considered, but will be introduced as context domains in Chapter 5.

Therefore the knowledge base needs to be structured so that it can supply the answer to the following two questions:

- Return a list M that given an `objective` lists all the possible `model` to fulfil this objective.
- Given M returns a list A that includes all the `assumption` to be tested and which `model` they relate to.

The elements of the SKB are denoted as follows:

Definition 3.1. Elements in the Statistical Knowledge Base

- The set of *models*: $M = \{m_1, \ldots, m_K\}$ where $k = 1, \ldots, K$
- The set of *assumptions*: $A = \{a_1, \ldots, m_P\}$ where $p = 1, \ldots, P$
- The set of *objectives*: $O = \{o_1, \ldots, o_Q\}$ where $q = 1, \ldots, Q$

The following relationships are defined in the SKB:

Definition 3.2. Relations in the Statistical Knowledge Base

- $F \subseteq M \times O$ where if m_k fulfils objective o_q then $(m_k, o_q) \in F$
- $C \subseteq M \times A$ where if a_p is a critical assumption for m_k then $(a_p, m_k) \in C$
- $OBJ \subseteq O \times O$ where if o_r is an alternative objective to o_q then $(o_r, o_q) \in O$

An example of a knowledge base is in figure 3.3. In Figure 3.3 the relations are:

- $\{(m_{s1}, o_s), (m_{s2}, o_s), (m_{s3}, o_s), (m_{b1}, o_n), (m_{b2}, o_n)\} = F$
- $\{(m_{s1}, a_1), (m_{s2}, a_1), (m_{s3}, a_a), (m_{s2}, a_2), (m_{s3}, a_3), (m_{b1}, a_{10}), (m_{b2}, \neg a_{10})\} = C$
- $\{(o_s, o_n)\} = OBJ$

3.4 Summary

In this chapter I have given an overview of the process of statistical model selection, the inputs that are required and a high level overview of the proposed approach. I have introduced an example and explained how analysis of this data given a specific hypothesis would be performed, including the steps required in order to select the suitable statistical models to apply. I have then highlighted the elements that are required in order to select a statistical model and discussed their relations. In Section 3.3 of this chapter I introduced a format for the knowledge base that will hold the required information in support of the methodology. In the next chapter I will focus on the proposed argumentation scheme, its associated critical questions and their interaction with the knowledge base.

Chapter 4

Argumentation Schemes for Statistical Model Selection

Chapter 3 presented an overview of the core elements of the approach I am proposing in support of the automation of statistical model selection through the use of argumentation schemes. In Section 4.1 I will introduce the following original contributions: an argument scheme to instantiate arguments in support of using model m_i to achieve objective o_j and its associated critical questions. I will delve into the details of the proposed argumentation scheme for statistical model selection, the critical questions and their interaction with the SKB. In Section 4.2 I will make use of the example introduced in Section 3.1.1 to illustrate the argument scheme and critical questions introduced in Section 4.1.

4.1 Argumentation Schemes

The arguments in support of the use of specific models (or model families) will be instantiated through argumentation schemes. There will be an argumentation scheme for model to consider on the grounds of meeting the objective. When this scheme is instantiated it will leverage the additional argumentation schemes as part of the critical questions in order to identify potential under-cutters, rebuttals or undermines.

A note regarding the notation: The notation I will use makes use of capital italics for the values in the argumentation scheme definitions (e.g.M for model) and I will make use of lower case italics for the values used for the instantiations of the argument schemes (e.g. m_i for a specific model).

Definition 4.1. (AS1): Argument Scheme for model to consider on grounds of achieving the objective - PM(R, O, M)

Premise - O is the objective of the research question R

Premise - M is a model able to analyse O

$\therefore M$ is suitable to answer R

The premises for this argumentation scheme (Definition 4.1) are statements that are verified against the SKB, given the specific o_q and r that the clinician is interested in. In order to evaluate the argument in support of the use of m_i it needs to be subject to critical questions.

Critical Questions

The proposed argument scheme for model to consider on grounds of achieving the objective (PM) (Definition 4.1) is to be subject to critical questions. These will either be used to test the assumptions of the scheme (such as CQ2) for potential undercuts or to highlight exceptions (CQ1) or rebuttals. The critical questions identified are:

- CQ1: Are there alternative ways of answering R? [This could lead to using another objective as this would also support the same analysis objective through a different set of models]

- CQ2: Do any of the critical assumptions for M fail to hold?

The critical questions are themselves defined in terms of argumentation schemes. The additional argumentation schemes for the critical questions are:

- CQ1: Argument Scheme for alternative objective (AO) - this will produce rebuttal arguments.

- CQ2: Argument Scheme for failed assumptions (CA) - this will produce arguments against the use of a model if any of its critical assumptions fail. This undercuts the arguments generated by the argument scheme for models to considers on the grounds of the objective.

The Argument Schemes are instantiated on the data and the clinician to generate a set of arguments in support of or against the use of each of the models that satisfy the objective of the research question.

Definition 4.2. (AS2): CQ1: Argument for alternative objective: AO(R, O, O_{alt})

- O_{alt} is an alternative objective to answer R
- M is a model able to analyse O_{alt}

$\therefore$ M is suitable to answer R

The premises for this argumentation scheme CQ1 (Definition 4.2) are statements to be extracted from the SKB once the initial R and O are known.

Definition 4.3. (AS3): CQ2: Argument against the use of a Model for failed critical assumptions : CA((M)

- Model M is suitable to answer R
- A is a critical assumption for M
- A does not hold

$\therefore$ $\neg M$ (M is not a model to be considered)

The first two premises for CQ2 (Definition 4.3) are statements that are extracted from the SKB for a given model to be considered m_i, and the last premise is validated either by performing an analysis on the data or by querying the clinician. In future the option of applying an argument from expert opinion for the assumptions that are confirmed by the clinician will be explored.

The instantiation of all of these argumentation schemes (Definitions: 4.1 4.2 4.3) will produce a set of arguments in support of or against the use of a specific model. These arguments are the set of arguments that make up the argumentation framework of relevant arguments to the research question and data at hand. The attack relations within this argumentation framework are relevant and derived from the desire to run only the most appropriate models, this implies that a decision to use one model (with arguments in support of its use) negates the use of other models with arguments supporting their use in the argumentation framework.

If it is acceptable to run all the models that have an argument in the argument framework that supports their use then this argumentation framework does not have any attack relations per se. If it is desirable to run only the most suitable models the relative strength or preference of each argument in support of the use of a model should be considered. In order to pick the most suitable of the models the context domains (non critical assumptions) and additional preferences over the models are introduced.

The process of accounting for these preference relations and leveraging them within the argumentation process will be covered in Chapter 5 of this thesis.

The relation between the different argumentation schemes and critical questions presented in this section is in Figure 4.1. The arrows refer to the relation between the argumentation schemes, AS2 (Definition 4.2) and AS3 (Definition 4.3) are instantiated as critical questions to AS1 (Definition 4.1). AS1 is instantiated to generate a set of arguments in support of the use of models that can achieve the objective of the research question, the critical questions then are instantiated as argument schemes. In cases where there is no alternative objective O_{alt} then there is no need to re-instantiate AS3 as no new arguments are generated in support of the use of alternative models.

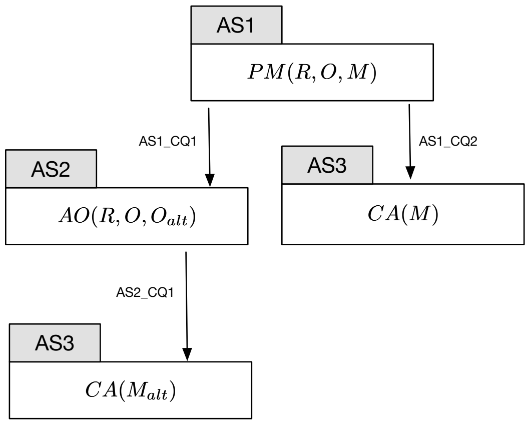


FIGURE 4.1: Argumentation schemes connected via their critical questions

4.2 ovarian example

The proposed argumentation scheme and critical questions will be applied to the example introduced in Section 3.1.1. The research question or hypothesis to be tested is Hypothesis 1 that looked at evaluating whether there was a difference in survival between treatment 1 and treatment 2. The hypothesis to be tested is equivalent to

r the research question to be used in the instantiation of the argumentation schemes introduced in Section 3.1.1. The target or dependent variable is survival of the patient, which is of type *time to event* and is therefore corresponding to an analysis objective of *survival analysis* o_s. The SKB contains the relations between the `objective` and the `model`, as well as the latter and the existence of any alternative objectives o_a.

The target for this hypothesis (Hypothesis 1) is a *time to event* attribute, called `futime` and as such the objective of the research question is o_s. The section of the SKB relevant to this research question relate to two objectives. These are also visually illustrated in Figure 3.3:

- Survival Analysis o_s which includes the following models: m_{s1} *Kaplan-meier*, m_{s2} *Proportional Hazards* and m_{s3} *weibull*
- Table Analysis o_n which includes the following models: m_{b1} χ^2 (Chi squared) , m_{b2} Fisher's Exact

The critical assumptions relevant to this example are:

- a_1 non informative censoring
- a_2 proportional hazards
- a_3 *weibull* distribution
- a_{10} table cell minimum > 5

The contents of the SKB:

- $\{(m_{s1}, o_s), (m_{s2}, o_s), (m_{s3}, o_s), (m_{b1}, o_n), (m_{b2}, o_n)\} = F$
- $\{(m_{s1}, a_1), (m_{s2}, a_1), (m_{s3}, a_a), (m_{s2}, a_2), (m_{s3}, a_3), (m_{b1}, a_{10}), (m_{b2}, \neg a_{10})\} = C$

- $\{(o_s, o_n)\} = OBJ$

The following facts about the assumptions are known (the outcomes of the relevant assumptions that have been tested on the data or by asking the clinicians):

- $\{a_1, a_2, \neg a_3, \neg a_{10}\}$.

Instantiations of AS1 (Definition 4.1) in this example leads to the following:

Arg_1 **Argument Scheme for model to consider on grounds of achieving the objective o_s : PM(o_s, r, m_{s1})**

Premise - o_s is the objective of the research question r

Premise - m_{s1} is able to analyse o_s

$\therefore$ - m_{s1} is suitable to answer r

Arg_2 **Argument Scheme for model to consider on grounds of achieving the objective o_s : PM(o_s, r, m_{s2})**

Premise - o_s is the objective of the research question r

Premise - m_{s2} is able to analyse o_s

$\therefore$ - m_{s2} is suitable to answer r

Arg_3 **Argument Scheme for model to consider on grounds of achieving the objective o_s : PM(o_s, r, m_{s3})**

Premise - o_s is the objective of the research question r

Premise - m_{s3} is able to analyse o_s

$\therefore$ - m_{s3} is suitable to answer r

Three arguments have been instantiated, these now need to be subject to the two critical questions. Instantiating CQ1 using AS2 (Definition 4.2) generates the following argument:

Arg'_4 **Argument for alternative objective: AO(o_s, r)**

- o_s is the objective of research question r
- o_n is an alternative objective to answer r
- m_{b1} is able to analyse o_n calling AS1:$PM(o_n, r, m_{b1})$

$\therefore$ - m_{b1} is suitable to answer r

Arg_4 **Argument for alternative objective: AO(o_s, r)**

- o_s is the objective of research question r
- o_n is an alternative objective to answer r
- m_{b2} is able to analyse o_n calling $PM(o_n, r, m_{b2})$

$\therefore$ - m_{b2} is suitable to answer r

- Arg_1 : PM(o_s, r, m_{s1}): m_{s1}
- Arg_2: PM(o_s, r, m_{s2}): m_{s2}
- Arg_3: PM(o_s, r, m_{s3}): m_{s3}
- Arg'_4: AO(o_s, r): m_{b1}
- Arg_4: AO(o_s, r): m_{b2}

Instantiating CQ2 using AS3 (Definition 4.3):

Arg_7 **Argument against the use of a Model for failed critical assumption: CA(**(m_{s3})

- Model m_{s3} achieves objective o_s
- a_3 is a critical assumption for m_{s3}
- a_3 does not hold

$\therefore$- m_{s3} is not a model to be considered

Arg_8 **Argument against the use of a Model for failed critical assumption: CA(**(m_{b1})

- Model m_{b2} achieves objective o_s
- a_{10} is a critical assumption for m_{b2}
- a_{10} does not hold

$\therefore$ - m_{b1} is not a model to be considered

The instantiation of the Argument Scheme AS3 (Definition 4.3) has generated two undercuts to two of the arguments in favour of the use of the respective models.

- Arg_1 : PM(o_s, r, m_{s1}): m_{s1}
- Arg_2: PM(o_s, r, m_{s2}): m_{s2}
- Arg_3: PM(o_s, r, m_{s3}): m_{s3}
- Arg'_4: AO(o_s, r): m_{b1}
- Arg_4: AO(o_s, r): m_{b2}
- Arg_7: CA$(((m_{s3}) : \neg m_{s3}$
- Arg_8: CA$(((m_{b2}) : \neg m_{b1}$

As argumentation schemes AS3 (Definition 4.3) can generate an undercut to the argumentation scheme AS1 (Definition 4.1) for possible model (PM) then this undercuts some arguments in support of models from the list of ones that could be applied in this case study. The resulting list is:

- Arg_1 : PM(o_s, r, m_{s1}): m_{s1}
- Arg_2: PM(o_s, r, m_{s2}): m_{s2}
- Arg_4 : PM(o_n, r, m_{b2}): m_{b2}

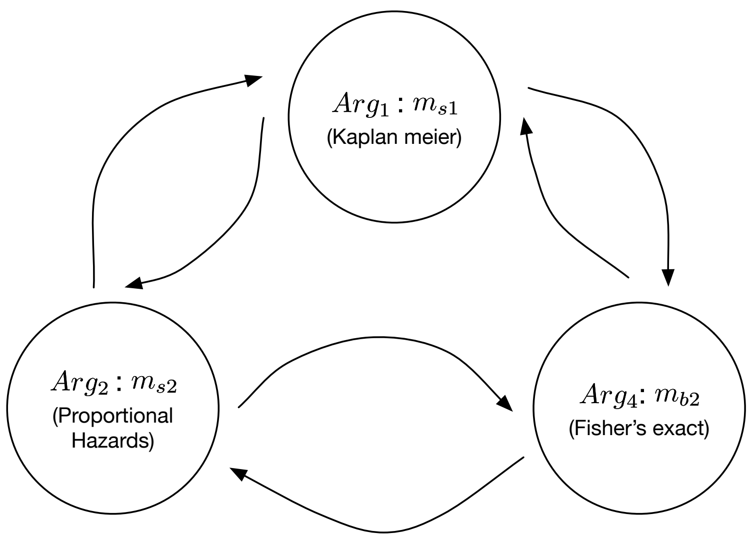


FIGURE 4.2: Argumentation Framework for the `ovarian` example with symmetrical attacks

In this situation we have an argumentation framework containing three arguments each supporting the use of a different model, this is illustrated in Figure 4.2. If the overall aim is to recommend one model or aim to refine the list of models to apply then this proposed methodology will need to be extended to include and account for the information that can assist in arguing what contexts are relevant to the situation and leverage them in order to recommend the most suitable model(s).

4.3 Summary

In this chapter I have introduced a novel argumentation scheme that enables the instantiation of arguments in support of the use of a model m_i on grounds of achieving the objective of the research question. The argumentation scheme leverages the knowledge base. The critical questions are defined to ensure that arguments in support of additional analysis approaches are considered, if relevant. The critical questions also validate whether any of the models' critical assumptions fail and therefore an argument against the use of those models is instantiated. The latter argument will undercut the argument instantiated by the argumentation scheme.

By applying the proposed methodology given a research question and data will result in a set of arguments each in support of the use of a model. Some aspects of the contributions presented in this chapter are published in [68].

The next step will be to reason with these arguments and to incorporate additional information in order to provide a recommendation as to the most suitable model or models to use (amongst all those that have an argument in favour of their use) given the circumstances. This will be addressed in the next chapter.

Chapter 5

Preferences

The aim of this chapter is to introduce preferences into the statistical model selection approach involving argument schemes and critical questions proposed in Chapter 4. Section 5.1 covers the first original contribution of this chapter and focuses on how the preferences are derived in the context of statistical model selection. Section 5.2 will review the main approaches used when applying preferences to argumentation, the methods will be described with an emphasis on the applicability to our scenario. In Section 5.3 my proposed approach will be explained, including how the SKB is extended to support preferences. The `ovarian` example introduced in Section 3.1.1 is used in Section 5.4 to illustrate the approach proposed in this Chapter. The original contributions in this chapter are detailed in section 5.1 and 5.3.

5.1 Preferences for Statistical Model Selection

In this section I will characterise and discuss the different source of preference orders that are relevant to statistical model selection. A preference expressed in the context of statistical model selection refers to an order of priority between models. I will use the notation $\succ$ to denote a preference relation. For example if we consider a set of only three models $\{m_1, m_2, m_3\}$ then if $m_1 \succ m_2$ indicates that m_1 is preferred to m_2. If we have a set of models $\{m_1, \ldots, m_n\}$ then a preference order $Pref \subseteq M \times M$ where $pref$ on these models would be of the form: $pref = \{m_1 \succ \cdots \succ m_i \succ \cdots \succ m_n\}$ where $i \in \{1, \ldots, n\}$. In the context of statistical model selection there will be multiple preference orders over the same set of models therefore I will also make use of the notation $pref_j = \{m_1 \succ \cdots \succ m_i \succ \cdots \succ m_n\}$ where $i \in \{1, \ldots, n\}$ and $j = 1, \ldots, J$. The latter index j refers to the different preference orders over the models. Preference orders could be empty, in which case all models would be equally preferred, and not all models need to be included in each preference order.

Preference orders over models will arise from different sources in the context of statistical model selection. One source for preference orders is the statistical theory underpinning each model and dictating which models perform better when certain conditions are present in the data or the research question. For example, certain types of model are more resilient to particular features in the data, e.g. censoring or the proportion of case data lost to follow up, whereas others tend to become unreliable in such circumstances. Here, the presence of a particular feature causes a preference ordering over statistical models to arise. This relationship between a feature and an associated preference ordering is a matter of statistical knowledge. The presence of the feature may be determined by applying a test on the data or needs to be elicited from domain knowledge. In what follows, such preferences are called feature-based preferences.

A second source of preference orders is derived from model intent. There are different reasons for building a model when answering a research question. McBurney [54] explores the different purposes or reasons why a model can be used. In the context of statistical analysis the two most common intents for building a model on data are the need to predict or the need to explain (understand) the data. This is also covered in detail in [72] where Shmueli tackles the distinction between explanatory modelling and predictive modelling in detail and the implications these have on the choice of model to use. The definition of a good model will differ depending on whether we are looking for explanatory or predictive power, and this will reflect itself in an order of preference between models that can achieve a specific analytic objective. This preference order between models will change depending on the intent (purpose) of the analysis. In what follows, such preferences are called intent-based preferences.

Finally, there may be preference orders that are derived from the Clinicians themselves. This could be due to the fact they are more familiar with a model, or that the literature they reference most makes use of a particular model. These preference orders can arise when more than one clinician is involved in an analysis and are an important factor within the decision making process. In what follows, such preferences are called domain-based preferences.

The source of the preference order can be mapped to a priority or importance when leveraging the preferences to find the model most suitable to the situation. Feature based preferences are generally more important in determining a models' relative suitability than intent based preferences or domain based preferences. Furthermore the intent based preferences are more important than domain based preferences. This can be represented as an order of importance over the different preference orders that are relevant to each type of preference source.

5.2 Applying preferences in argumentation

The approach adopted for statistical model selection makes use of argumentation, various approaches have been devised to incorporate preferences into argumentation models to help decide what arguments to accept or reject and facilitate decision making. I have described these in Section 2.7. In this section I will initially expand on some of the methodologies described in Section 2.7 and illustrate how they would be applicable to statistical model selection using the example introduced in Section 3.1.1.

There are a number of approaches proposed to accomplish the task of taking preferences into account as part of the argumentation process. However not all are applicable to statistical model selection. The main approaches I am going to explore in depth in this section are the methodologies which have some applicability within the context of statistical model selection.

5.2.1 Preferences in Statistical Model Selection example

I will be using the `ovarian` example introduced in Section 3.1.1 to illustrate how these approaches relate to statistical model selection. When instantiating the argument schemes (Definitions 4.1, 4.2, 4.3) given the objective of testing Hypothesis 1 then the resulting argumentation framework contained the following arguments:

- Arg_1 : PM(o_s, r, m_{s1}): m_{s1}
- Arg_2: PM(o_s, r, m_{s2}): m_{s2}
- Arg_4 : PM(o_n, r, m_{b2}): m_{b2}

These arguments when represented as an argumentation framework $\langle Arg, \mathcal{R} \rangle$ are $R = \{(m_{s1}, m_{s2}), (m_{s1}, m_{b2}), (m_{s2}, m_{s1}), (m_{s2}, m_{b2}), (m_{b2}, m_{s1}), (m_{b2}, m_{s2})\}$. Where (m_{s1}, m_{s2}) represents an attack of the argument supporting the use of m_{s1} on the argument in support of the use of m_{s2}. The argumentation framework is illustrated in Figure 5.1 where the symmetrical attacks are illustrated.

The preference orders over the models, each of which results from a different context or source are:

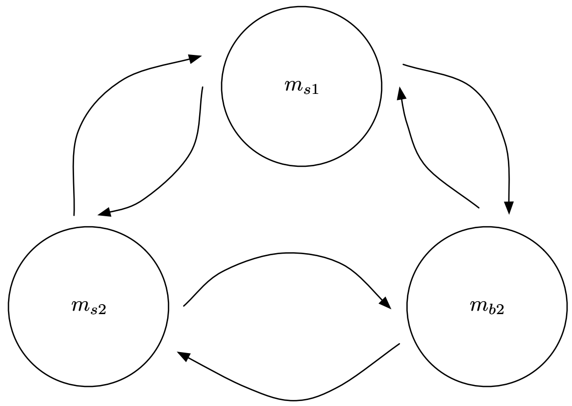


FIGURE 5.1: Arguments for three possible models attacking each other

- cd_1: $pref_1 = m_{s1} \succ m_{s2} \succ m_{b2}$
- cd_2: $pref_2 = m_{b2} \succ m_{s2} \succ m_{s1}$
- cd_3: $pref_3 = m_{s1} \succ \{m_{s2}, m_{b2}\}$

There is an order of importance I to each of these contexts (cd_i or sources of preference orders) within the SKB. This list will be different for each different objective, as different objectives will need different sets of models. In this example the order of context domains is: $I = \{cd_1 \succ cd_2 \succ cd_3\}$, where the most important order is cd_1. The preference order $pref_1$ that results from context domain cd_1 results in the following preferences: $\{m_{s1} \succ m_{s2}, m_{s1} \succ m_{b2}, m_{s2} \succ m_{b2}\}$. This ordering can also be represented as $\{(m_{s1}, m_{s2}), (m_{s2}, m_{b2}), (m_{s2}, m_{b2})\}$ where (m_{s1}, m_{s2}) is equivalent to $m_{s1} \succ m_{s2}$.

Table 5.1 contains all the relevant sets of preferences ordered by the context and importance. These preference orders contain conflict. If one were to only consider cd_3 then m_{s1} would be preferred, however if one only considered cd_2 then m_{b2} would be preferred.

Importance	Source	Preference order
cd_1	Censoring: mild	$pref_1 = \{(m_{s1}, m_{s2}), (m_{s1}, m_{b2}), (m_{s2}, m_{b2})\}$
cd_2	Model Intent:explain	$pref_2 = \{(m_{b2}, m_{s2}), (m_{b2}, m_{s1}), (m_{s2}, m_{s1})\}$
cd_3	Clinician Preference	$pref_3 = \{(m_{s1}, m_{s2}), (m_{s1}, m_{b2})\}$

Table 5.1: Preference orders and relative importance for the `ovarian` example

5.2.2 Preference Based Argumentation Frameworks

In the context of argumentation the use of preferences in decision support offers a way of representing the strength or priority of an argument. Preferences can also be used to quantify the quality or uncertainty underlying an argument. In their article Amgoud *et al.* [4] introduce the concept of a preference based argumentation framework (PAF) as a way to leverage preference relations into an argumentation framework, as a refinement of Dung's acceptability calculus [31]. The authors note that a weakness in Dung's definition of acceptability is it disregards the quality of the argument. Their proposal for preference based argumentation combines the preference relations between arguments with the defeasibility relations, in other words preferences between arguments are considered at the same time as attacks between argument to determine if the attack is valid.

The authors differentiate between individual acceptability and joint acceptability of an argument[4]. The former is related to the existence of direct defenders, whilst the latter explores the existence of sets of arguments that can be accepted as they defend themselves against defeaters. The authors propose a general preference based argumentation framework (PAF) where an argument is acceptable if: it is not defeated, if it defends itself against its defeaters or it is defended by other arguments.

A PAF is defined in [4] as $\langle Arg, \mathcal{R}, Pref \rangle$ and $\geqslant^{Pref}$ denotes the strict ordering associated with $pref$. $A, B \in Arg$ then B attacks A if and only if $B\mathcal{R}A$ and not $(A \geqslant^{Pref} B)$. The following categories are then defined with a PAF:

- S_a are the acceptable arguments of the AF
- $S_r = \{A \in Arg \mid \exists\, B \in S_a$ such that $B\mathcal{R}A$ and not $(A \geqslant^{Pref} B)\}$
- $S_s = Arg\ (S_a \cup S_r)$ in abeyance.

The set of acceptable arguments will define both S_r and S_s, so in order to construct S_a the authors propose two notions of defence using the preference relations:

- $A, B \in Arg$ such that $B\mathcal{R}A$. A defends itself against B if and only if $(A \geqslant^{Pref} B)$. A defends itself against B as it is preferred to it.
- $C_{R,Pref}$ contains the set of arguments that defend themselves against their defeaters.
- $S \subseteq Arg$. An argument A is defended by S if and only if $\forall\, B \in A$ if $B\mathcal{R}A$ and not $(A \geqslant^{Pref} B)$ then $\exists\, C \in S$ such that $C\mathcal{R}B$ and not$(B \geqslant^{Pref} C)$.

The authors show that the acceptable arguments are those which defend themselves from their defeaters ($C_{R,Pref}$) and also arguments which are defended by the arguments of $C_{R,Pref}$. The authors claim that in order to know the given status of any argument in the framework it is not necessary to calculate all the sets of arguments : S_a, S_r, S_s

The authors define a strict defence in a PAF $\langle A, \mathcal{R}, Pref\rangle$ as follows: $A \in Arg$ and $S \subseteq Arg$. A is strictly defended by S if and only if $\forall\, B \in A$ such that $B\mathcal{R}A$ then $\exists\, C \in S$ such that C disqualifies B. The definition of disqualification is as follows: An argument C disqualifies another argument B if and only if $C\mathcal{R}B$ and not $B\mathcal{R}C$. In order to verify if an argument is acceptable then only its strict defenders need to be taken into account.

In a more recent paper Amgoud *et al.* [6] show that three proposed extensions of the Dung framework may lead to unintended results, and propose a new preference based argumentation framework that ensures sound results. The authors define two basic requirements of any PAF: extensions should be conflict free with respect to the attack relation, and in case preferences are not available then the PAF extensions are the same as Dung'e extensions for an AF. The new approach proposed takes inputs from three elements: Arg a set of arguments, an attack relation $\mathcal{R}$ and a partial or total pre-order $\geq$. It returns extensions that are subsets of Arg which satisfy the following:

- Conflict-freeness: if ε is an extension of $\langle Arg, \mathcal{R}, \geq \rangle$, then ε is conflict free with respect to $\mathcal{R}$
- Generalisation: If $(\nexists A, B \in Arg)$ such that $(A, B) \in (R)$ and $(B, A) \in >$, then any extension $\langle Arg, \mathcal{R}, \geq \rangle$ is also an extension of Dung's framework $\langle Arg, \mathcal{R} \rangle$ and vice versa.

In a further refinement of PAFs [7] Amgoud *et al.* discuss the dual role of preferences within an argumentation framework. They explain that preferences are used to repair the attack relation between arguments (handling critical attacks), and to refine the evaluation of arguments (refine the results of a framework). The authors claim that only the first of these roles has been addressed within the existing literature on the topic. The authors demonstrate that existing models that repair the attack relation with preferences do not perform well and in certain situations may return counter intuitive results. The authors then propose a new abstract framework to accomplish both aims of the use of preferences.

Attacks that conflict with the preference relation between arguments are called critical if $\langle Arg, \mathcal{R} \rangle$ and $\geqslant \subseteq Arg \times Arg$, an attack $B\mathcal{R}A$ is critical if and only if $A > B$

The authors explain that methods proposed whereby a critical attack is removed from the framework can lead to non conflict free extensions for the framework. In the case of symmetric attacks this can be shown not to be the case, as the conflict between the arguments in question still remains, and they wont be both in any resulting extensions. However the requirement for symmetric attacks is very restrictive. The authors further suggest that the removal of any attack relations results in information being lost.

Amgoud *et al.* propose that in case of critical attacks then the direction of the attack is reversed, not removed from the framework. This guarantees that arguments involved in a critical attack cannot be together in the same extension. Dung's semantics are then applied to the repaired (modified) graph. Their modification is defined as follows: The extensions T under a given semantics are the extensions of the AF $\langle Arg, \mathcal{R} \rangle$ called repaired frameworks under the same semantics with: $R_r = \{(A, B) \mid (A, B) \in \mathcal{R}$ and not $(B > A)\} \cup \{(B, A) \in \mathcal{R}$ and $B > A\}$.

In this proposed repaired framework for critical attacks the arrow is inverted and the preferences take precedence over attacks. Preferences do not take precedence over attacks when the attacks are not critical. It follows that if a PAF has no critical attacks then the repaired PAF is equivalent to a PAF. This approach delivers always conflict free extensions. The authors further show that in cases of symmetric attacks their proposed methodology leads to the same result as removing the attack. Furthermore the authors show that when the attack relation is symmetric the extensions of a PAF are a subset of those of its basic AF, the preferences help filter the extensions to the best ones.

The authors also discuss the role of preferences in refining the argumentation frameworks, they refer to this as refinement. Refinement is the use of preferences on the extensions resulting from the repaired framework, to select the preferred extensions. In order to do so the authors use the concept of either a democratic or elitist relation. These are defined as follows:

- Democratic relation: Let Δ be a set of objects and $\geqslant \subseteq \Delta \times \Delta$ be a partial pre-order. For $X, X' \subseteq \Delta$, $X \geqslant_d X'$ if and only if: $\forall x' \in X' \; X. \exists x \in X \; X'$ such that $x > x'$.

- Elitist relation: Let Δ be a set of objects and $\geqslant \subseteq \Delta \times \Delta$ be a partial pre-order. For $X, X' \subseteq \Delta$, $X \geqslant_e X'$ if and only if: $\forall x \in X \; X', \exists x' \in X' \; X$ such that $x > x'$

Intuitively under *democratic* d for each item in the less preferred set there is a better element in the preferred set, and under *elitist* e for each item in the more preferred set there is an item in the other set that is less preferred than it.

The authors propose a new definition of a rich preference based argumentation framework as an abstract model that extends Dung with preferences between arguments and integrates both roles of preferences (Repairing and Refining). The preferences are accounted for in a two step process firstly the attack relations are repaired by computing R_r, to generate an argumentation framework $\langle Arg, \mathcal{R}_r \rangle$. The preferences are then used to refine the extensions computed from $\langle Arg, \mathcal{R}_r \rangle$. The latter can be achieved using for example either the democratic or elitist principle for comparing extensions.

PAF and the example

In order to illustrate the use of PAFs in the context of preferences in statistical model selection I will be using the `ovarian` example and the preference orders introduced in Section 5.2.1. When applying PAFs to this example we encounter a major restriction in that only one set of preferences would be considered. Assuming only one set of non contradicting preferences cd_1 from Table 5.1: $m_{s1} \succ m_{s2} \succ m_{b2}$ then:

- PAF is $\langle Arg, \mathcal{R}, Pref \rangle$ where $Arg = \{m_{s1}, m_{s2}, m_{b2}\}$ and $Pref = m_{s1} \succ m_{s2} \succ m_{b2}$.

- $S_a = \{m_{s1}\}, S_r = \{m_{s2}, m_{b2}\}$
- S_a is conflict free.

Furthermore the repaired framework corresponding to the PAF involves reversing the direction of the attack from m_{s2} to m_{ms1}, m_{b2} to m_{s2}, m_{b2} to m_{s1} resulting in Figure 5.2.

FIGURE 5.2: Argumentation Framework and corresponding repaired preference argumentation framework for the example

5.2.3 Contextual Preferences argumentation frameworks

An argumentation framework based on contextual preferences (CPAF) is defined by Amgoud *et al.* [5] as a tuple $\langle Args, \mathcal{R}, C, \succ, Pref_1, \ldots, Pref_n \rangle$. A is a set of arguments, $\mathcal{R}$ is a binary relation representing the defeat relationship between the arguments, $C = \{c_1, \ldots, c_n\}$ is a set of contexts, $\succ$ is a complete preordering on $C \times C$, $Pref_i$ is a (partial or complete) preordering on $A \times A$ issued from the context c_i.

The second step in the argumentation process, after the construction of arguments and counter arguments, is the selection of the most acceptable arguments. In order to select these arguments within a CPAF the authors suggest three solutions: (1) aggregating the different preference relations, (2) changing the definitions of individual and joint defence, (3) aggregating the sets of acceptable arguments.

The first solution aggregates the preference relations by keeping the preferences expressed in the most privileged context and among the remaining contexts keep only the preference relations that don't contradict the ones already kept. The result of this aggregation is Pref= $\prod_n$ such that:

- $T_1 = C$
- $\prod_1 = \{(A, B) \in Pref_i$ such that $\forall c_j \in T_1\ \{c_i\}, c_i \succ c_j\}$
- $T_{k+1} = T_k\ \{c_i\}$
- $\prod_{k+1} = \prod \cup \{(A, B) \in Pref_i, c_i \in T_{k+1}$, such that $(B, A) \notin \prod_k$ and $\forall c_j \in T_{k+1}\ \{c_j\}, c_i \succ c_j\}$

Once the aggregation is done the acceptable arguments are found by computing the set S_a of the PAF $\langle Arg, \mathcal{R}, Pref \rangle$.

The second solution involves computing a new set of acceptable arguments, by changing the definition of both individual and joint defence so that all the preferences are taken into account. A defeated argument must be preferred to its defeaters in a context which takes precedence over all contexts where the opposite preference is expressed in order to self defend.

Formally if $A, B \in Arg$ such that $A\mathcal{R}B$. B defends itself against A if and only if $\exists\, c_i \in C$ such that $B \geqslant^{Pref_i} A$ and $\forall\, c_j$ such that $A \geqslant^{Pref_j} B$ then $c_i \succ c_j$. $C_{R,\succ}$is the set of arguments defending themsleves against their defeaters.

Let $S \subseteq Arg$ and $A \in Arg$. S defends A if and only if $\forall\, B\mathcal{R}A$ and A does not defend itself against B then $\exists\, C$ such that $C\mathcal{R}B$ and B does not defend intself against C.

The set of acceptable arguments of the framework $\langle Arg, \mathcal{R}, C, \succ, Pref_1, \ldots, Pref_n \rangle$ denoted by S_{a2} is: $S_{a2} = \cup F^{i \geq 0}(\varnothing) = C_{R,\succ} \cup [\cup F^{i \geq 1}(C_{R,\succ})]$

The third solution involves computing the acceptable arguments for each of the PAFs $\langle Arg, \mathcal{R}, Pref_1\rangle, \ldots \langle Arg, \mathcal{R}, Pref_n\rangle$ and denote these as $S_1, \ldots, S_n$. The acceptable arguments of this CPAF are intially those in S_i such that c_i is the most privileged context. Then the next most privileged context c_j is selected and arguments from S_j are kept only if they are not defeated by an argument already in S_i. Formally the set of acceptable arguments of the framework is $S_{a3} = \prod_n$ such that:

- $T_1 = C$
- $\prod_1 = \{A \in S_i$ such that $\forall\, c_j \neq c_i, c_i \succ c_j\}$
- $T_{k+1} = T_k\ \{c_i\}$
- $\prod_{k+1} = \prod_k \cup \{A \in S_i, c_i \in T_{k+1}$ such that $\forall\, c_j \in T_{k+1}\ \{c_i\}, c_i \succ c_j$ and $\nexists B \in \prod_k$ such that $B\mathcal{R}A$ and $B \geqslant^{Pref_l} A$ with $c_l \notin T_{k+1}\}$

The authors state that $S_{a1} = S_{a2} = S_{a3}$. Intuitively the order in which the acceptable arguments are computed and aggregated yields the same set of acceptable arguments.

CPAF and the example

In order to implement an approach similar to the one proposed in [5] using the `ovarian` example, the different sets of preference orders will be defined as $cd_1, ...cd_n$ where each cd_i contains a complete or partial pre order of the models $pref_i = \{(m_{s1}, m_{s2}), (m_{b2}, m_{s2})\}$. (m_{s1}, m_{s2}) indicates that m_{s1} is preferred to m_{s2}. There will be an importance order on the different preference orders such that $\{cd_1 \succ cd_2 \succ \cdots \succ cd_n\}$, this is equivalent to $\{Pref_1 \succ Pref_2 \succ \cdots \succ Pref_n\}$. All of the preference orders and their relative importance are in Table 5.1. I will apply the three approaches proposed in [5] to combining the preference orders across the different contexts, while taking into account their relative importance or privilege.

The first method proposed involves creating a $Pref$ from $Pref_1, \ldots, Pref_n$ by keeping all the preferences expressed in their most important (or privileged context). In our example above this would result in a preference order that matches the one derived from c_1. $Pref = \{m_{s1} \succ m_{s2} \succ m_{b2}\}$. It would follow that as there are arguments in support of the use of m_{s2}, m_{s2}, m_{b2} then applying this preference order would result in recommending the use of m_{s1} over the others and the reason for this preference would be cd_1.

The second approach proposed involves computing a new set of acceptable arguments by changing the definition of defence to take into account the preferences. In order to self defend a defeated argument must be preferred to its defeaters in a context which is more important, than any of the other contexts where the opposite preference is expressed. In our example the arguments in support of m_{s1}, m_{s2}, m_{b2} attack each other. Applying the latter definition of defence the only argument that successfully self defends is m_{s1}.

The third approach proposed involves computing the set of acceptable arguments $S_1, \ldots, S_n$ for each of the the frameworks: $< Args, \mathcal{R}, Pref_1 >, \ldots, < Args, \mathcal{R}, Pref_n >$. Where $Args$ are the arguments, and $\mathcal{R}$ the attack relations between them. In our example we would compute S_1, S_2, S_3 by applying in turn the different preferences cd_1, cd_2, cd_3. $S_1 = \{m_{s1}\}$, $S_2 = \{m_{b2}\}$ and $S_3 = \{m_{s1}\}$. These sets of arguments will be aggregated keeping in the acceptable set only arguments that are not defeated by arguments present in a more important context. Applying the latter to the example results in m_{s1} being preferred.

When applied to our situation this example has resulted in the same model being recommended under each of the three solutions. The process for arriving to a recommended model under each of the the three solutions has differing levels of abstraction and complexity. The first solution is most intuitive and could be used as part of the justification of the model recommended to the end user.

5.2.4 Value argumentation frameworks

One approach makes use of a value assigned to each decision. Although there are many measures of model fit or prediction accuracy that can be used to measure a model, the choice of the most suitable model for the given data and research question may lead to a model that does not perform as well on these metrics as a model that was not suitable. Therefore in our case there is no relevant method of assigning values when selecting the most appropriate model therefore Value Argumentation Frameworks (VAF) [16] is not a suitable approach in this case.

5.2.5 Extended Argumentation frameworks

So far in the chapter the methods that have been included have all had a common theme in that preferences were expressed as orders over the arguments. A different approach is proposed by Modgil in[58] where he introduces the concepts of an extended argumentation framework that includes meta level arguments to enable arguing about the preferences. An extended argumentation framework (EAF) includes a second attack relation that ranges from arguments to attacks on the attacks.

Formally an EAF is defined as a tuple $\langle Arg, \mathcal{R}, \mathcal{D} \rangle$ such that Arg is a set of arguments and:

- $\mathcal{R} \subseteq Arg \times Arg$
- $\mathcal{D} \subseteq Arg \times \mathcal{R}$
- if $(X, (A, B)), (X'(B, A)) \in \mathcal{D}$ then $(X, X'), (X', X) \in \mathcal{R}$

Alternative notations are suggested by [58]:

- $A \rightharpoonup B$ is equivalent to $(A, B) \in \mathcal{R}$ A symmetric attack would be written as $A \leftrightharpoons B$.
- An attack X on the attack of A on B is written as: $X \twoheadrightarrow (A \rightharpoonup B)$ which is equivalent to $(X, (A, B)) \in \mathcal{D}$

In a extended argumentation framework the preferences are claimed as arguments. If A attacks B then A defeats B only if in the context of the set of arguments committed to, there is no argument claiming that B is preferred to A. Formally the concept of a strict defeat is defined as: Given an EAF $\langle Arg, \nabla, \mathcal{D}\rangle$ and a subset $S \subseteq Arg$ then A defeats$'_s$ B if and only if $(A, B) \in \mathcal{R}$ and $\neg\exists C$ such that $(C, (A, B)) \in \mathcal{D}$. Intuitively A strictly defeats B if A attacks B and there is no attack on the preference of A over B. If A defeats$_s B$ and B does not defeat$_s A$ then A strictly defeats$_s B$. The latter can be noted as $A \rightarrow^S B$.

Within an EAF $\langle Arg, \mathcal{R}, \mathcal{D}\rangle$ a set $S \subseteq Arg$ is conflict free if and only if $\forall A, B \in S$ if $(A, B) \in \mathcal{R}$ then $(B, A) \notin \mathcal{R}$ and $\exists C \in S$ such that $(C, (A, B)) \in \mathcal{D}$. This means that if $A, B \in S$ and A attacks B then S is conflict free only if B does not attack A and there is a C that attacks the attack from A to B.

In order to define an acceptable argument within an EAF the concept of a reinstatement set is to be defined first. $R_S = \{X_1 \rightarrow^S Y_1, \ldots, X_n \rightarrow^S Y_n\}$ is a reinstatement set for $C \rightarrow^S B$ if and only if:

- $C \rightarrow^S B \in R_S$
- for $i = 1, \ldots, X_i \in S$,
- $\forall X \rightarrow^S Y \in R_S$, $\forall Y'$ such that $(Y', (X, Y)) \in \mathcal{D}$, there is a $X' \rightarrow^S Y' \in R_S$.

$A \in Arg$ is acceptable with respect to $S \subseteq Arg$, if and only if: $\forall B$ such that $B \rightarrow^S A$, there is a $C \in S$ such that $C \rightarrow^S B$ and there is a reinstatement set for $C \rightarrow^S B$. The admissible, preferred, complete and stable extensions of an EAF are defined the same way as for Dung frameworks.

EAF and the example

In order to implement an extended argumentation framework using the arguments generated from the argument scheme from the `ovarian` example introduced in Section 5.2.1 the preference orders will be represented as meta-level arguments. For example if a preference order states that $m_{s1} \succ m_{s2}$ then the related meta-level argument generated from it is $P1 \in \mathcal{D}$. This can be represented as $(P1, (m_{s2}, m_{s1}))$ or $P1 \twoheadrightarrow (m_{s2} \rightharpoonup m_{s1})$. All of these preference arguments can be derived in EAF notation from Table 5.1, eight arguments $Pj \in D$ would be extracted from Table 5.1.

- From cd_1:
  - $P_{cd11} \twoheadrightarrow (m_{s2} \rightharpoonup m_{s1})$
  - $P_{cd12} \twoheadrightarrow (m_{b2} \rightharpoonup m_{s2})$
  - $P_{cd13} \twoheadrightarrow (m_{b2} \rightharpoonup m_{s1})$
- From cd_2:
  - $P_{cd21} \twoheadrightarrow (m_{s2} \rightharpoonup m_{b2})$
  - $P_{cd22} \twoheadrightarrow (m_{s1} \rightharpoonup m_{s2})$
  - $P_{cd23} \twoheadrightarrow (m_{s1} \rightharpoonup m_{b2})$
- From cd_3:
  - $P_{cd31} \twoheadrightarrow (m_{s2} \rightharpoonup m_{s1})$

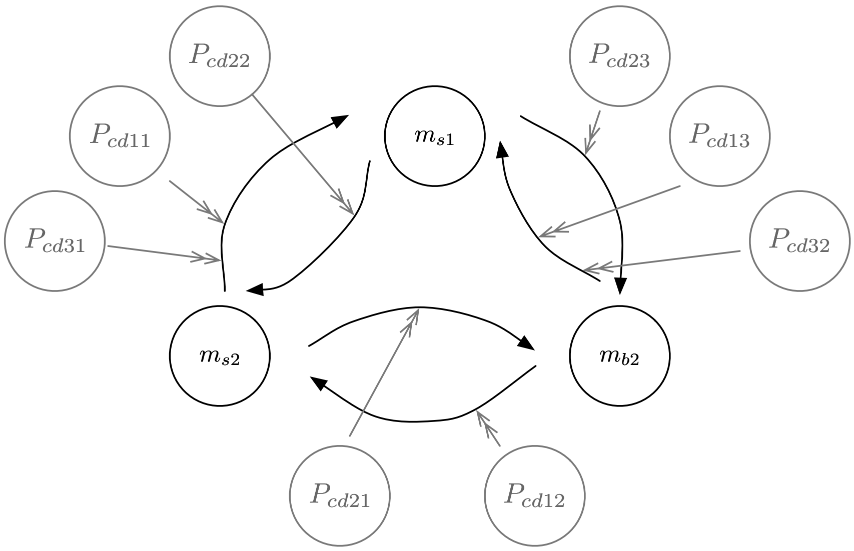


FIGURE 5.3: Overlaying all the meta level preference arguments on the `ovarian` example

- $P_{cd32} \twoheadrightarrow (m_{b2} \rightharpoonup m_{s1})$

These arguments in $\mathcal{R}$ when represented as an argumentation framework $\langle Arg, \mathcal{R} \rangle$ are $R = \{(m_{s1}, m_{s2}), (m_{s1}, m_{b2}), (m_{s2}, m_{s1}), (m_{s2}, m_{b2}), (m_{b2}, m_{s1}), (m_{b2}, m_{s2})\}$.

Where (m_{s1}, m_{s2}) represents an attack of the argument supporting the use of m_{s1} on the argument in support of the use of m_{s2}

The meta-level arguments in $\mathcal{D}$ are superimposed onto the existing attack relations R that exist between the three possible models. Figure 5.3. shows the resulting EAF when all of the meta level arguments resulting from the different preference orders are applied simultaneously. In such an argumentation framework all the conflicts between arguments are attacked by meta level arguments as such the resulting argumentation framework is devoid of conflict. This does not assist in selecting the most suitable model.

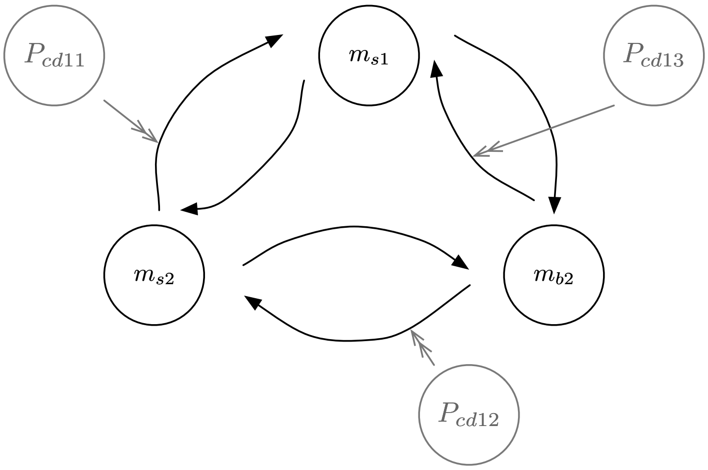


FIGURE 5.4: Extended Argumentation Framework for `ovarian` including only meta level arguments from one context domain

The EAF resulting from the example introduced in section 5.2.1 does not enable us to recommend an extension that contains arguments in support of one or more most suitable models, however its representation is very visual and offers potential to justify the decision through the use of the arguments. Applying the order of the contexts to the EAF would facilitate the existence of extensions containing arguments in support of the use of the most suitable models. When the meta level arguments from only one context (such as the most important one) are applied to the argumentation framework in Figure 5.3, the resulting EAF has a preferred extension that includes one argument in support for one model. Figure 5.4 shows the same EAF when only the most important context domain is applied. Figure 5.5 shows the remaining attacks between arguments, the ones attacked by the meta level arguments are in dotted lines. The resulting argumentation framework's preferred extension is such that $\{m_{s1}\}$ is acceptable with respect to it.

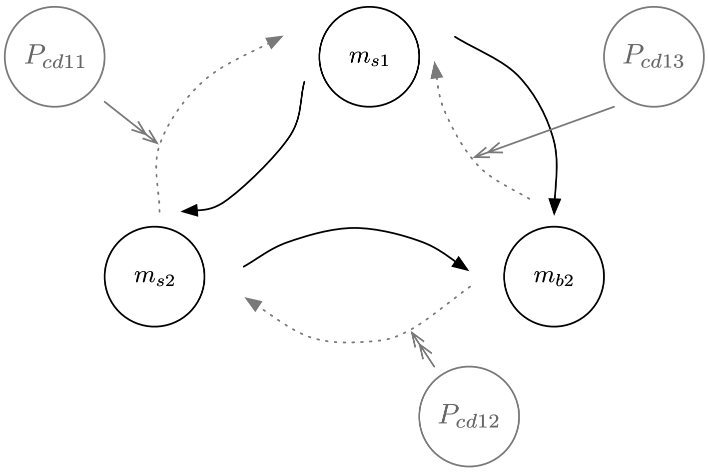


FIGURE 5.5: Extended Argumentation Framework for `ovarian` including only meta level arguments from one context domain, and remaining attack relations

5.3 Proposed Argumentation Model

In the previous section I illustrated different approaches to the use of preferences in argumentation frameworks. PAFs [4] were able to consider one set of preferences only, CPAFs [5] were able to cater for multiple preference orders however these were aggregated prior to their application to the attack relations in the Argumentation Framework. EAFs [58] provide the possibility of arguing about conflicting preferences as well as conflicts between arguments, as the preferences are represented as meta-level arguments. Furthermore in EAFs attacks between arguments and attacks on attacks from the preferences can be visualised and reasoned with simultaneously. These are the reasons why I chose to use EAFs to underpin my proposed method.

The proposed method in this thesis for leveraging preferences as part of statistical model selection relies on two components: an extension of the SKB and the extension of the argumentation framework to an extended argumentation framework (EAF) with context domain preference orders.

5.3.1 Extending the SKB to include preferences

To incorporate preferences into the approach, the statistical knowledge base (SKB) introduced in Chapter 3 is extended with the following:

Definition 5.1. Extended Statistical Knowledge Base

- A set of context domains $CD = \{CD_1, \ldots, CD_H\}$.
- A set of totally ordered sets of performance measures $P = \{P_1, \ldots, P_H\}$. Each P_h contains a set of measures $p_{h1} \prec \ldots \prec p_{hj}$ by means of which a model's performance is assessed in a specific context.

- A set of performance functions $PF = \{PF_1, \ldots, PF_H\}$, such that each $PF_i : CD_i \times M \mapsto P_i$. The mapping of the model to the performance measure is dependent on the context domain. CD_h
- Optionally an order of importance for the context domains can be defined where I is the ordered set of context domains. The order determines the relative importance of the context domain.

By applying Definitions 5.1 to the example introduced in Section 5.2.1 and using cd_1 from Table 5.1 then the definitions above would lead to the following:

- Assuming there is only one context domain cd_1 then $CD = \{cd_1\}$
- The performance measures for cd_1 are $P_{cd_1} = \{p_1, p_2, p_3\}$, where $p_1 \succ p_2 \succ p_3$.
- Then PF_1 maps models $\{m_{s1}, m_{s2}, m_{b2}\}$ to the applicable ordered performance measures $\{p_1, p_2, p_3\}$

Table 5.2 contains the definitions concerning $cd1_1, cd1_2$ and $cd1_3$ which are the context domains related to censoring (absent, mild or heavy). For example a model such as *Kaplan-meier* is mildly affected by light censoring and strongly affected by heavy censoring.

Table 5.3 contains the mappings for the two contexts that are related to intent based preferences $cd2_1$ and $cd2_2$.

5.3.2 Reasoning with Arguments and Preferences for Statistical Model Selection

To construct an argumentation model based on the extended statistical knowledge base, first the set of contexts $\widehat{CD} \subseteq CD_1 \cup \ldots \cup CD_H$ for the problem at hand must be

context domain	model m	Performance measure p
$cd1_1$ censoring absent	m_{s1} KM	p_1 = unaffected
	m_{s2} PH	p_1 = unaffected
	m_{b2} $Fisher's$	p_1 = unaffected
$cd1_2$ censoring light	m_{s1} KM	p_2 = mildly affected
	m_{s2} PH	p_1 = unaffected
	m_{b2} $Fisher's$	p_3 =strongly affected
$cd1_3$ censoring heavy	m_{s1} KM	p_3 = strongly affected
	m_{s2} PH	p_1 =unaffected
	m_{b2} $Fisher's$	p_3 = strongly affected

Table 5.2: $cd1_1$, $cd1_2$ and $cd1_3$ performance function mapping - subset of relevance to `ovarian` example

context domain	model m	Performance measure p
$cd2_1$ predict	m_{s1} KM	p_3 avoid
	m_{S2} PH	p_1 suitable
	m_{b2} $Fisher's$	p_3 avoid
$cd2_2$ explain	m_{s1} KM	p_1 suitable
	m_{s2} PH	p_1 suitable
	m_{b2} $Fisher's$	p_2 neutral

Table 5.3: $cd2_1$ and $cd2_2$ performance function mapping for model intent - subset of relevance to `ovarian` example

established. $\widehat{CD}$ contains the subset of contexts taken from all the context domains in CD. Whether a context is relevant to a problem is derived by applying a test on the data, elicited from the domain expert/clinician or elicited from the research question. Where identification of the context is not straightforward, the contexts in CD provide hooks (conclusions) for further arguments about the appropriate statistical model.

Let $\langle Arg, \mathcal{R} \rangle$ be an argumentation framework generated using the methods described in Chapters 3 & 4. Such a model can now be extended to an EAF $\langle Arg, \mathcal{R}, \mathcal{D} \rangle$ by defining:

Definition 5.2. Generating the Extended Argumentation Framework using the Extended Statistical Knowledge Base

- $\forall m_x, m_y \in Arg$, $(m_x, m_y) \in \mathcal{R}$ and $(m_y, m_x) \in \mathcal{R}$. m_x, m_y are arguments generated by instantiating the argument schemes and critical questions (Definitions 4.1, 4.2, 4.3) in support of the models m_x & m_y respectively.

- $\forall CD_h \in \widehat{CD}$ if $p_{cdh}(m_x) \prec p_{cdh}(m_y)$ there is a meta level argument $PA_{cdh_{xy}} \in \mathcal{D}$ such that $(PA_{cdh_{xy}}, (m_x, m_y))$. Intuitively, an attack relationship $PA_{cdh_{xy}} \twoheadrightarrow (m_x \rightharpoonup m_y)$ is added for each attack of a model m_x by a model m_y where a context (CD_h) justifies a preference of m_y over m_x.

Optionally there may be a preference order I over the context domains, $CDM \subseteq CD \times CD$. Intuitively a context domain cd_i may be of higher importance than a domain cd_j if the former is derived from statistical theory and the latter is clinician's preference.

The preferred models will be the ones which are supported by an argument that is acceptable with respect to the set of preference arguments from the context domain in question. If the arguments in support of the use of more than one model are acceptable to the set of preference arguments generated by one context domain, another context domain (the next one in order of importance) can also be used.

5.4 Representing the `ovarian` example as an EAF with Context Domains

In this section I will use the `ovarian` example used in Section 3.1.1. In Chapter 3 the instantiation of the argumentation scheme (Definition 4.1) and critical questions proposed in this thesis (Definition 4.2 and 4.3) have generated an argumentation framework $\langle Arg, \mathcal{R} \rangle$ where:

- $Arg = \{Arg_1 : m_{s1}, Arg_2 : m_{s2}, Arg_4 : m_{b2}\}$

context domain	model m	Performance measure p
$cd1_2$ light censoring	m_{s1} KM	p_2 = mildly affected
	m_{s2} PH	p_1 = unaffected
	m_{b2} $Fisher's$	p_3 =strongly affected

TABLE 5.4: $cd1_2$ performance function mapping relevant to `ovarian` example

- $\mathcal{R} = \{(Arg_1 : m_{s1}, Arg_2 : m_{s2}), (Arg_1 : m_{s1}, Arg_4 : m_{b2}), (Arg_2 : m_{s2}, Arg_4 : m_{b2}), (Arg_2 : m_{s2}, Arg_1 : m_{s1}), (Arg_4 : m_{b2}, Arg_1 : m_{s1}), (Arg_4 : m_{b2}, Arg_2 : m_{s2})\}$

In order to generate the EAF for this scenario then the following additional inputs are required in order to generate the meta level arguments $\mathcal{D}$:

- $\widehat{CD} = \{cd1_2, cd2_2, cd3\}$ where $cd1_2$ in this case corresponds to *light censoring*, $cd2_2$ is *model intent* and $cd3$ is clinician preference.
- $I = \{cd1_2 \succ cd2_2 \succ cd3\}$

Context Domain $cd1$'s mapping is available in Table 5.2 for *absent, light, heavy* censoring. The `ovarian` data is light censored (the proportion of censored patients is 54% which is mild). The relevant slice of Table 5.2 $cd1_2$ is presented in Table 5.4.

Applying definitions 5.2 then the following set of meta level arguments are generated:

- $(PA_{cd1-12}, (m_{s1}, m_{s2}))$ as $p_{cd1}(m_{s1}) \prec p_{cd1}(m_{s2})$
- $(PA_{cd1-13}, (m_{b2}, m_{s2}))$ as $p_{cd1}(m_{b2}) \prec p_{cd1}(m_{s2})$
- $(PA_{cd1-23}, (m_{b2}, m_{s1}))$ as $p_{cd1}(m_{b2}) \prec p_{cd1}(m_{s1})$
- where $PA_{cd1-12}, PA_{cd1-13}, PA_{cd1-23} \in \mathcal{D}$

context domain	model m	Performance measure p
$cd2_2$ explain	m_{s1} KM	p_1 suitable
	m_{s2} PH	p_1 suitable
	m_{b2} $Fisher's$	p_2 neutral

TABLE 5.5: $cd2_2$ performance function mapping for model intent ='explain' relevant to `ovarian` example

The second context domain of relevance in this example is $cd2_2$. In this case the intent of the analysis is to explore the hypothesis (not to predict) therefore an additional set of meta level arguments are generated from Table 5.3 where $cd2_2 =' explore'$. The relevant aspects of Table 5.3 are replicated in Table 5.5.

Applying definition 5.2 then the following set of meta level arguments are generated:

- $(PA_{cd2-13}, (m_{b2}, m_{s1}))$ as $p_{cd2}(m_{b2}) \prec p_{cd2}(m_{s1})$
- $(PA_{cd2-12}, (m_{b2}, m_{s2}))$ as $p_{cd2}(m_{b2}) \prec p_{cd2}(m_{s2})$
- where $PA_{cd2-13}, PA_{cd2-12} \in \mathcal{D}$

Finally there is a clinician expressed preference, this will be $cd3$ and the clinician has expressed a preference for m_{s1}. This also generated a set of meta level arguments:

- $(PA_{cd3-13}, (m_{s2}, m_{s1}))$ as $p_{cd3}(m_{s2}) \prec p_{cd3}(m_{s1})$
- $(PA_{cd3-12}, (m_{b2}, m_{s1}))$ as $p_{cd3}(m_{b2}) \prec p_{cd3}(m_{s1})$
- where $PA_{cd3-13}, PA_{cd3-12} \in \mathcal{D}$

There are now seven meta level arguments in $\mathcal{D}$. FIgure 5.7 illustrates the extended argumentation framework generated and including the meta level arguments (preference arguments PAs) from each of the three context domains.

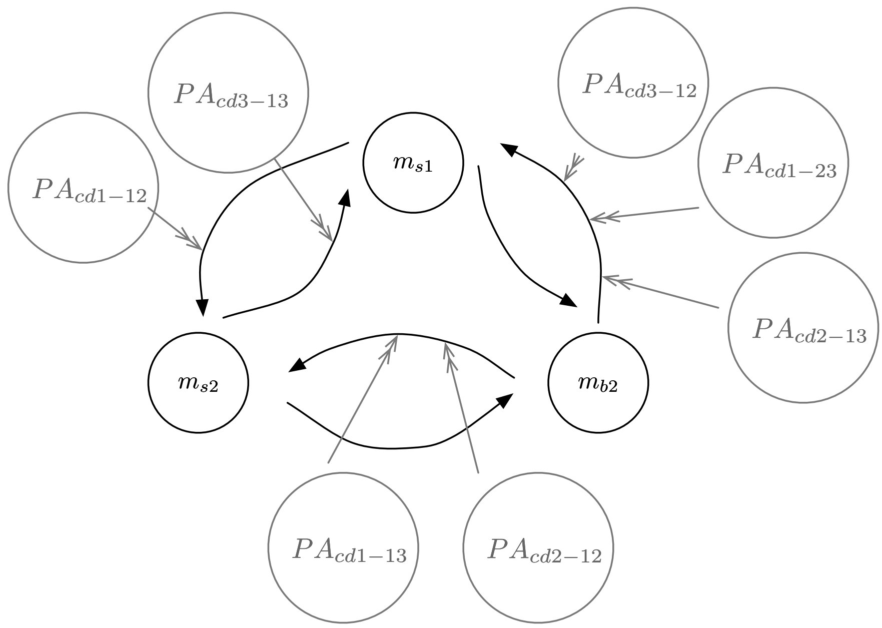


FIGURE 5.6: Extended Argumentation Framework for `ovarian` including preference arguments from all three context domains

If the order of importance of context domains is applied and only the meta-level arguments generated from $cd1_2$ are applied there is a preferred extension which contains the preference arguments from $cd1_2$ and the argument in support of $\{m_{s2}\}$. $\{m_{s2}\}$ is therefore acceptable with respect to the arguments in the EAF considered.

The preferences can be resolved in order to determine the recommended model by initially only considering the preference arguments from the most important context domain ($cd1_2$). When only applying the preferences derived from $cd1_2$ then m_{s2} is acceptable with respect to $S'_{cd1} = \{PA_{cd1-12}, PA_{cd1-13}, PA_{cd1-23}\}$, is the only argument that is not strictly defeated and as such this would be the recommended model to be used. In this EAF, the justification to its choice over m_{s1} and m_{b2} is given by the context domain used in order to resolve this: $cd1_2$. In this case the recommendation of m_{s2} over the other models is explained by it being preferred under conditions of mild censoring.

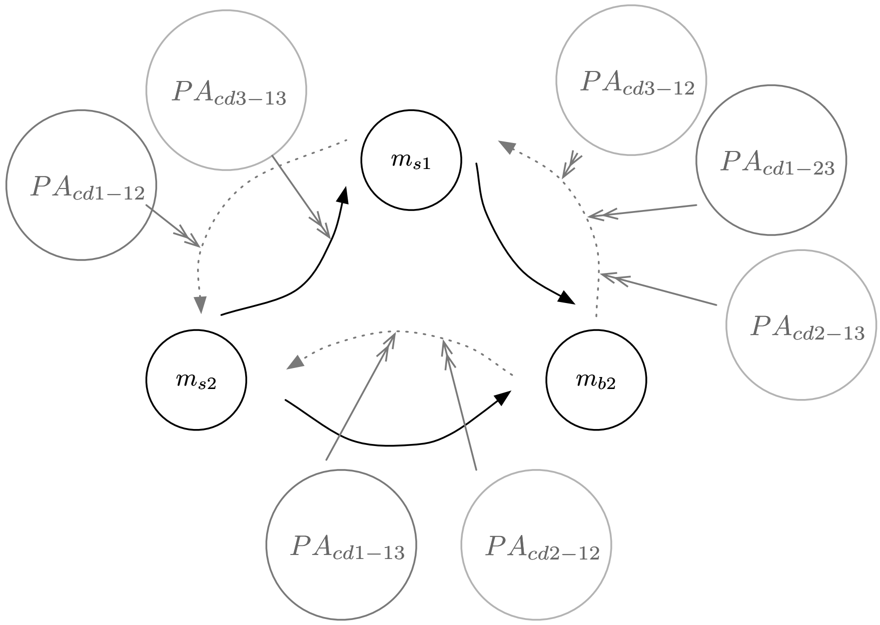


FIGURE 5.7: Extended Argumentation Framework for `ovarian` including preference arguments from $cd1_2$ only

If we assume that the order over the context domains in the same example is not given, then the admissable extensions for the EAF can be computed considering the preference arguments derived from each cd_i in turn. The resulting extensions would be: $S_1 = \{m_{s2}\}$, $S_2 = \{m_{s2}\}$, $S_3 = \{m_{s1}\}$ for $cd1_2$, $cd2_2$ and $cd3$ respectively. In other words model m_{s1} would only be selected in a situation where the preferences of the clinician ($cd3$) are prioritised over all other contexts.

5.5 Summary

The approached proposed in this chapter supports the statistical model selection process by enabling conflicting preference orderings to be accounted for and reasoned with in order to recommend the most suitable models. Aspects of the original contributions made in this chapter have been published in [69]. The differing preference orders are incorporated into statistical model selection through generating meta level arguments, extending the argumentation framework into an EAF and with the support of an extended statistical knowledge base. This approach can also take into account the relative importance of the different preference context domains, if this is applicable to the situation. This approach further provides a justification for any recommendation made through the use of context domain importance, or offers a method of acknowledging the context domain that is prioritised if a specific model is selected (when an importance order is not relevant or not available).

Chapter 6

Formalising the Argument Scheme , Critical Questions and Extended Statistical Knowledge Base

In this chapter I will formalise the original contributions outlined in this thesis using Z notation. The formalisation of the system will help clarify the operations of the proposed methodology, enable representation in a computational system and enable reasoning over and about the system. Furthermore the formalisation will also aid in the development of a prototype implementation. Section 6.1 introduces Z notation and articulates its use in similar situations. Section 6.2 implements the argument scheme, critical questions and SKB as articulated in Chapter 4. Section 6.3 provides an overview of the roles and relations between the different schemas proposed in Section 6.2. Section

6.4 extends the Z specification introduced in Section 6.2 to include preferences through the extension of the SKB.

6.1 Z notation

Z notation [74] is based on elementary components such as set theory and first order predicate logic. There are examples of the use of Z notation to formalise multi agent systems ([52, 29, 56]). Luck *et al.* [52] use Z to provide an accessible and formal account of agent systems. The authors use Z as it is sufficiently expressive to allow a consistent unified structured account of a system and its associated operations and it is deemed suitable to facilitate implementation. Z notation was also used by Miller *et al.* [56] as a basis for an extension of Z aimed at modelling software agents in a multi agent environment.

An additional implementation in a multi agent setting that makes use of Z is provided by D'inverno *et al.* [29]. The authors use Z to provide an abstract formal model of an idealised *dMARS* (distributed Multi Agent Reasoning System). The authors justify the use of Z as it enables designs of systems to be formally developed, whilst allowing for the systematic reduction of these specifications to implementation. The authors also describe Z as having the desirable property of being accessible and extremely expressive allowing for consistent unified and structured accounts of systems. Furthermore the large array of books and cases studies (academic and industry) is also cited.

The expressiveness and accessibility of Z notation coupled with the need to facilitate a prototype implementation for the contributions proposed within this thesis supported my decision to use it as a basis for the formalisation of the contributions. Additional auxiliary background material on Z notation is provided in Appendix A Section A.1.

6.2 Formalising the Argument Scheme, Critical Questions and SKB for Statistical Model Selection in Z Notation

I will now start to build the specifications required to represent the methodology and related original contributions outlined in Chapters 3 and 4 in Z notation. There are two aspects of the proposed methodology that need to be addressed: the knowledge base and the argumentation schemes. The former will be specified first in order for the latter to be able to leverage the concepts introduced to specify the inference.

Initially the basic types required to define the elements of the SKB (Definition 3.1) need to be introduced:

[MODEL] - the set of all possible models

[OBJECTIVE] - the set of all possible objectives

[ASSUMPTION] - the set of all possible assumptions

There is also a need to strengthen the specification by setting up variables to account for potential input errors. As example of such a situation is if a user attempts to find models for an undefined objective, or to find assumptions for a model that is not defined. In order to do so I will first define a type:

$$REPORT ::= ok \mid already\_known \mid not\_known$$

$$\begin{array}{|l} \hline [Success] \\ result! : REPORT \\ \hline result! = ok \\ \hline \end{array}$$

This schema can be then used in conjunction with another schema such as [*FindModels*] to flag the situation where the input model is not known to the system.

$$FindModels \wedge Success$$

The state space for the proposed specification is:

$$\begin{array}{|l} \hline [StatisticalKnowledgeBase] \\ known : \mathbb{P}\, MODEL \\ achieves : MODEL \leftrightarrow OBJECTIVE \\ requires : MODEL \leftrightarrow ASSUMPTION \\ \hline \end{array}$$

The relations between the different elements (Definition 3.2) to be defined by the *[StatisticalKnowledgeBase]* schema are illustrated in Figure 6.1 where the light grey lines represent the *achieves* relation and the dashed black lines represent the *requires* relation.

Initially the Statistical Knowledge Base will be empty

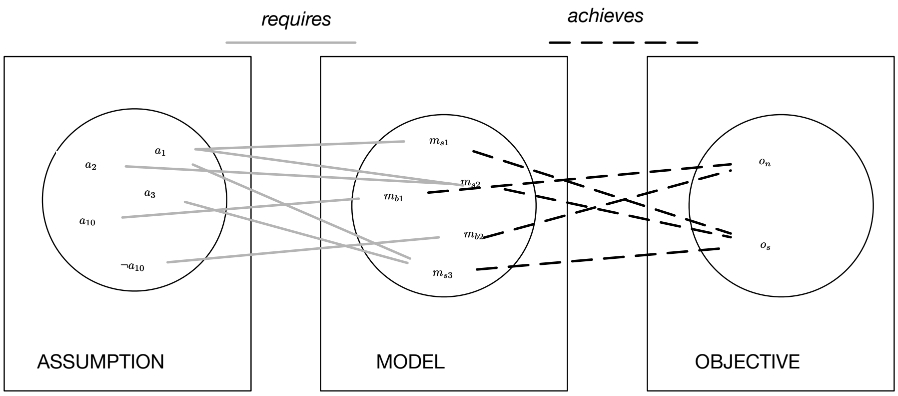


FIGURE 6.1: The *requires* and the *achieves* relationship in the SKB

$$[IntitSKB]$$
$$StatisticaKnowledgeBase$$
$$known = \{\}$$

An example of the contents of the knowledge base used for the `ovarian` example introduced in Section 3.1.1 and illustrated in Figure 3.3 is:

$known = \{m_{s1}, m_{s2}, m_{s3}, m_{b1}, m_{b2}\}$

$achieves = \{m_{s1} \mapsto time\_to\_event, m_{s2} \mapsto time\_to\_event,$

$m_{s3} \mapsto time\_to\_event, m_{b1} \mapsto nominal$

$m_{b2} \mapsto nominal\}$

$requires = \{m_{s1} \mapsto a_1, m_{s2} \mapsto a_1, m_{s2} \mapsto a_2,$

$m_{s3} \mapsto a_1, m_{s3} \mapsto a_3, m_{b1} \mapsto a_{10}, m_{b2} \mapsto \neg a_{10}\}$

$\mathrm{dom}\, achieves = \{m_{s1}, m_{s2}, m_{s3}, m_{b1}, m_{b2}\}$

$\mathrm{ran}\, achieves = \{time\_to\_event, nominal\}$

$\mathrm{dom}\, requires = \{m_{s1}, m_{s2}, m_{s3}, m_{b1}, m_{b2}\}$

$\mathrm{ran}\, requires = \{a_1, a_2, a_3, a_{10}, \neg_{10}\}$

In Z notation the notion of a related pair is defined as a `maplet` $\mapsto$. The following $x \mapsto y \in F$ is equivalent to $(x, y) \in F$.

A new entry to *[StatisticalKnowledgeBase]* is added with the following:

$[AddModelObjective]$

$\Delta StatisticalKnowledgeBase$

$model? : MODEL$

$objective? : OBJECTIVE$

$known' = known \cup model?$

$achieves' = achieves \cup \{model? \mapsto objective?\}$

$[AddModelAssumption]$

$\Delta StatisticalKnowledgeBase$

$model? : MODEL$

$assumption? : ASSUMPTION$

$known' = known \cup model?$

$requires' = requires \cup \{model? \mapsto assumption?\}$

Note that the difference in approach in this formalisation means that there is no requirement that the model is not known, as more than one objective can be mapped to a known model. Furthermore the process of assigning an objective and assumption to a model is now split into two schemas. Given that a model can have more than one objective and more than one assumption then this also means that these schemas can be called multiple times for each model.

The instantiation of the argument scheme (AS1) (Definition 4.1) given an objective is in the form of the following enquiry schema:

$[FindModels]$

$\Xi StatisticalKnowledgeBase$

$objective? : OBJECTIVE$

$models! : \mathbb{P}\, MODEL$

$result! : REPORT$

$(objective? \in \operatorname{ran} achieves \wedge$

$models! = \{m : model \mid (objective? \mapsto m) \in achieves\}$

$\wedge\ result! = ok) \vee$

$(objective? \notin achieves \wedge$

$result! = not\_known)$

where the type REPORT will be defined to flag situations where the objective stated is not known.

$$REPORT ::= ok \mid not\_known \mid none$$

The result of instantiating *[FindModels]* is either a list of models and a confirmation that the results are OK, or a message reporting that the objective of the research question is not defined in the relation *achieves*.

Given the list of models, the critical questions need to be instantiated. *CQ1: Are there alternative ways of answering the research question?* In order to model this critical question an additional relation is to be introduced to the *[StatisticalKnowledgeBase]*.

This will be a relation between OBJECTIVEs.

$$alternative : OBJECTIVE \leftrightarrow OBJECTIVE$$

This relation can created through this schema:

$$[AlternativeObjective]$$

$$known\_objectives : \mathbb{P}\, OBJECTIVE$$

$$alternative : OBJECTIVE \leftrightarrow OBJECTIVE$$

$$known\_objectives = \operatorname{dom} alternative$$

In order to set this relation up the schema is initialised:

$$[InitAlternativeObjective]$$

$$AlternativeObjective$$

$$known\_objective = \operatorname{ran} achieves$$

Then, the schema to populate the relationship between alternative objectives is:

$$
\begin{array}{l}
[AddAlternativeObjective] \\
\hline
\Delta AlternativeObjective \\
objective1? : OBJECTIVE \\
objective2? : OBJECTIVE \\
result! : REPORT \\
\hline
(objective1? \in \operatorname{ran} achieves \land objective2? \in \operatorname{ran} achieves \\
\land\ alternative' = alternative \cup \{objective1? \mapsto objective2?\} \\
\land\ result! = ok) \\
\lor\ (\{objective1? \notin \operatorname{ran} achieve \lor objective2? \notin \operatorname{ran} achieves\} \\
\land\ result! = not\_known) \\
\hline
\end{array}
$$

The above schema will add a relation between one objective (o_1) and another objective (o_2) that can be used as an alternative analysis approach to it. The schema will only allow a relation to be added if both o_1 and o_2 are defined in the Statistical Knowledge Base [StatisticalKnowledgeBase].

Now the argument scheme AS2 in support of the first critical question (Definition 4.2) can be instantiated through the following schema:

$[FindAlternativeObjective]$

$\Xi AlternativeObjective$

$objective1? : OBJECTIVE$

$objectives2! : \mathbb{P}\, OBJECTIVE$

$result! : REPORT$

$(objective2! = \{o : objective \mid (objective1? \mapsto o) \in alternative\}$

$\wedge\ result! = ok) \vee (objective2! = \{\} \wedge result! = none)$

The second critical question's aim (Definition 4.3) is to check the critical assumptions for each of the models returned as part of *[FindModels]* (note that the latter will be instantiated for both the original objective and any additional alternative objectives resulting from *[FindAlternativeObjective]*)

$[FindAssumptions]$

$\Xi StatisticalKnowledgeBase$

$model? : MODEL$

$assumptions! : \mathbb{P}\, ASSUMPTION$

$result! : REPORT$

$(assumptions! = \{a : assumption \mid (a \mapsto model?) \in require\}$

$\wedge\ result! = ok)$

$\vee$

$(assumptions! = \{\} \wedge result! = none)$

The resulting list of assumptions will contain elements of the type: $\{a_1, a_2, \ldots, a_n\}$ and these can each then be validated against either the data or the clinician. This relation between assumptions and their type can also be formalised in a schema.

A new type can be defined to aid in the distinction between assumptions that are to be tested against the data and those that require eliciting a response from the clinician.

$$TYPE ::= query \mid user$$

The state space for this is defined as:

$[AssumptionType]$

$$assumption : ASSUMPTION$$
$$type : TYPE$$
$$assumption\_type : ASSUMPTION \leftrightarrow TYPE$$

The *[AssmptionType]* is initialised:

$[InitAssumptionType]$

$$AssumptionType$$

$$assumption\_type = \{\}$$

The relation between the assumptions and their types can be defined by adding the relation to the knowledge base:

$[AddAssumptionType]$

$\Delta AssumptionType$

$assumption? : ASSUMPTION$

$type? : TYPE$

$result! : REPORT$

$(assumption? \in requires \land$

$assumption\_type' = assumption\_type \cup \{assumption? \mapsto type?\}$

$\land\ result! = ok)$

$\lor$

$(assumption? \notin requires \land$

$assumption\_type' = assumption\_type$

$\land\ result! = not\_known)$

The critical question enquire schema is:

$[FindAssumptionType]$

$\Xi AssumptionType$

$assumption? : ASSUMPTION$

$type! : TYPE$

$result! : REPORT$

$(type! = (t : type \mid assumption? \mapsto t) \in assumption\_type \land$

$result! = ok)$

$\lor$

$(type! = (t : type \mid assumption? \mapsto t) \notin assumption\_type \land$

$result! = not\_known)$

6.3 Overview of the Z schemes proposed

In the previous section I articulated the Z notation specifications required to set up the SKB, to instantiate the argument scheme and critical questions. The specification proposed is made up of three different categories of schemas: schemas to initialise the knowledge base, schemas to populate the relations in the knowledge base and schemas to instantiate the argument schemes and critical questions.

The types defined and used by the schemas proposed in Section 6.2 are:

$REPORT ::= ok \mid not\_known \mid none$

$TYPE ::= query \mid user$

The following state spaces are defined:

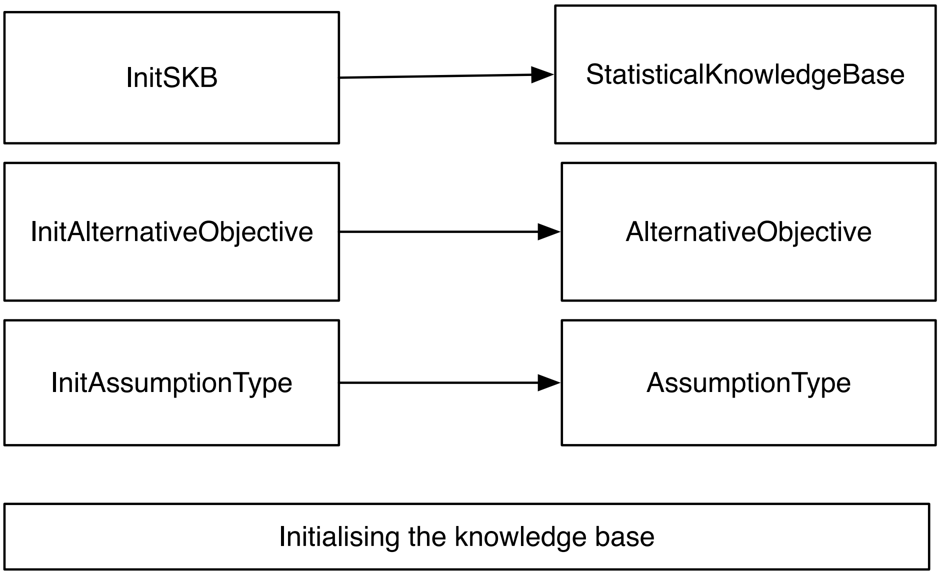


FIGURE 6.2: The Z schemas used to initialise the elements of the SKB on the left and the elements of the SKB on the right

- StatisticalKnowledgeBase
- AlternativeObjective
- AssumptionType

The following schemas are used to populate the relations within the knowledge base. The relations between these different schemas used to initialise the knowledge base is in Figure 6.2, for example the schema *[initAlternativeObjective]* is used to initialise the *[AlternativeObjective]*. Figure 6.3 illustrates the relations between the schemas to modify the knowledge base and the construct they are relevant to.

- InitSKB
- AddModelObjective
- AddModelAssumption

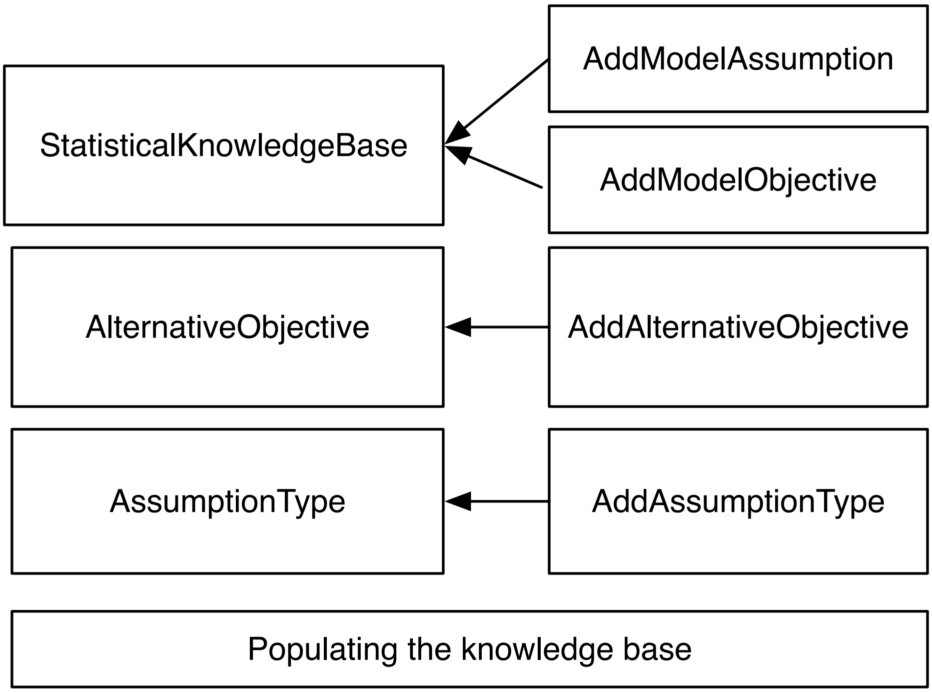


FIGURE 6.3: The Z schemas to populate the SKB on the right and their related element on the left

- InitAlternativeObjective
- AddAlternativeObjective
- InitAssumptionType
- AddAssumptionType

The following enquire schemas represent the argument scheme and its associated critical questions:

- FindModels
- FindAlternativeObjective

- FindAssumptions
- FindAssumptionType

In the next section the formalisation in Z notation will be expanded to include the extended knowledge base as proposed in Chapter 5.

6.4 Formalising the Extended SKB for Statistical Model Selection in Z Notation

In Chapter 5 I proposed an extension of the knowledge base aimed at incorporating and leveraging multiple sets of conflicting preference orders over the models in order to provide not just a set of possible models to apply given a data set and a research question but to refine this to a set of recommended models taking additional contextual factors into account through the use of an extended argumentation framework.

Given the performance measures are the source of the model orders it makes sense for these to be represented as a sequence. The position of each item is relevant as it will determine which performance measure value is better within the relevant context domain.

In order to expand the SKB two new types (Definition 5.1) will be introduced to the specification:

$$[CONTEXT\_DOMAIN]$$
$$[PERFORMANCE\_MEASURE]$$

where there are a set of predefined values that [PERFORMANCE_MEASURE] can take:

$$PERFORMANCE\_MEASURE ::= affected \mid unaffected \mid neutral$$
$$\mid suitable \mid avoid \mid clinician\_pref$$

The state space for this is defined as follows:

[ContextDomainBase]

$known\_domain : \mathbb{P}\, CONTEXT\_DOMAIN$

$p\_measure : \mathbb{P}\, PERFORMANCE\_MEASURE$

$model : \mathbb{P}\, MODEL$

$measured : CONTEXT\_DOMAIN \leftrightarrow PERFORMANCE\_MEASURE$

$relevant : CONTEXT\_DOMAIN \leftrightarrow MODEL$

$effect : PERFORMANCE\_MEASURE \leftrightarrow MODEL$

$known\_domain = \operatorname{dom} measured = \operatorname{dom} relevant$

The [ContextDomainSpace] is initialised:

[InitContextDomainBase]

$ContextDomainBase$

$known\_domain = \{\}$

In order to populate this knowledge base initially the context domains are defined:

$[AddContextDomain]$

$\Delta ContextDomainBase$

$domain? : CONTEXT\_DOMAIN$

$measures? : PERFORMANCE\_MEASURE$

$res! : REPORT$

$(domain \notin known\_domain \wedge$

$measured' = measured \cup \{domain? \mapsto measures?\} \wedge$

$res! = ok) \vee$

$(domain \in known\_domain \wedge$

$res! = already\_known)$

For each given context domain (defined with the previous two schemas) the models relevant to it can be added and mapped onto the existing defined performance measures. So a schema is to be defined that for a given model and context domain will map the model to a performance measure.

$[AddModelToContext]$

$\Delta ContextDomainBase$

$model? : MODEL$

$domain? : CONTEXT\_DOMAIN$

$res! : REPORT$

$(domain? \in known\_domain \wedge$

$relevant' = relevant \cup \{model? \mapsto domain?\} \wedge$

$effect' = effect \cup \{model? \mapsto measure\}$

$res! = ok)$

$\vee$

$(domain? \notin known\_domain \wedge$

$res! = not\_known)$

The *[AddModelToContext]* schema will take a model and context domain as inputs. Then it will map the effect to be assigned to the model from the list of ones defined for the specific context domain in question only. For example if we have a context domain for censoring $cd1_2$ as defined in Table 5.2 then the values would be:

$$domain = \{cd1_1\}$$

$$measures = \{p_1, p_2, p_3\}$$

where p_1 = *unaffected*, p_2 = *mildly affected* and p_3 = *strongly affected.*

The *AddContextDomain* schema would be used to define this context domain (assuming it has not been already defined). Then the *AddModelToContext* schema would be employed to map all the models the context domain is relevant to to the appropriate measures. In this example using context domain $cd1_2$ from Table 5.2 as an example:

$$effect = \{m_{s1} \mapsto p_2, m_{s2} \mapsto p_1, m_{b2} \mapsto p_3\}$$

Which results in the following preference between models resulting from $cd1_2$: $\{m_{s2} \succ m_{s1}, m_{s2} \succ m_{b2}, m_{s1} \succ m_{b2}\}$

The definition of the context domain that is derived from clinician preference would also be mapped using the Z schemas *[AddContextDomain]* and *[AddModelToContext]*. An example of such a context domain where there are only two performance measure $p_1 and p_2$ where $p_1 \succ p_2$ would lead the the following mapping if the clinician prefers the use of model m_s2:

$$effect = \{m_{s1} \mapsto p_2, m_{s2} \mapsto p_1, m_{b2} \mapsto p_2\}$$

Which results in the following preference between models resulting from the specified clinician preferences: $\{m_{s2} \succ m_{s1}, m_{s2} \succ m_{b2}\}$

6.5 Conclusion

In this chapter I articulated a formalisation of the extended statistical knowledge base, argument scheme and the critical questions based on the contributions presented in Chapters 3, 4 and 5 of this thesis.

The requirement for a formalisation using Z was derived from the need to document the system fully in preparation for a prototype implementation. This formalisation provided the schemas required to initialise such a system as well as the schemas for the argumentation scheme instantiations. The formalisation in Z notation provided one of the inputs for the development of the prototype. The use of the prototype to evaluate the contributions of this thesis is discussed in Chapter 7.

Chapter 7

Towards Evaluation

In this chapter I will cover different aspects of evaluation. In Section 7.1 I provide a review on how other argumentation based decision support methodologies have been evaluated. In Section 7.2 I propose a set of evaluation criteria relevant to the original contributions of this thesis. Within the scope of this thesis the evaluation is based on case studies which are elaborated in Section 7.3. The desired future evaluation tasks are also articulated in this chapter. Section 7.4 presents details of the prototype that was developed based on the original contributions of this thesis. The method of evaluation employed in scope of this thesis is the case studies in Section 7.3.

7.1 Evaluating Argumentation based Decision Support Systems

An important consideration as to how such a decision support system would be evaluated is by reviewing how other similar systems have been evaluated. The relevant argumentation based decision support systems have been reviewed in Section 2.6: EIRA, CARREL and DRAMA.

The ArguEIRA decision support system [42] flags anomalies in a patient's response to medication. The knowledge required by the system to flag these anomalies was derived from interviews with clinicians. The benefits of the system were assessed by three clinicians in the relevant clinical domain. The main focus of this evaluation was on the layout of the information reported back to the clinician when an anomaly was flagged. Further evaluation to assess whether the argumentation schemes used in EIRA can be ported to a different (non medical) domain was suggested.

The collaboration with clinicians in order to evaluate a proposed argumentation based decision support system was also key in CARREL [81]. Tolchinsky *et al.* evaluated the system on a set of examples, this was used to determine whether the argumentation schemes and their associated critical questions captured all of the required lines of reasoning. A specific aim of the evaluation was to ascertain that the template for the argument schemes was not too onerous but still captured all the required details.

A case study forms one of the key aspects of evaluation for the DRAMA agent [13]. The proposed system based on the DRAMA agent is not evaluated as such but is illustrated through the use of a running example. The evaluation criteria employed by Atkinson *et al.* in [13] initially focused on listing the proposed decision support's worthwhile features and stating its potential, but is further explored and evaluated in [9]. In [9] the evaluation of the method is achieved through the following criteria: comprehensiveness of the model, flexibility of the model when used in different domains and realistic representation of real life arguments. These criteria were selected as representing important elements in line with the aim of the thesis [9].

The approach to evaluation selected in [9] is one that could be applied to my thesis, however the resulting criteria would be different. As my thesis' remit is within clinical decision support then all of the scenarios are to be limited to this domain, however these can span different objectives within the domain. Therefore I propose that the initial the evaluation criteria for my proposed methodology consist of case studies based on clinical data and research questions from clinicians.

The use of case studies to evaluate and assess a proposed method is also covered by McBurney *et al.* [55]. In [55] the authors present a formal framework for deliberation dialogues, grounded in a theory of deliberative reasoning from the philosophy of argumentation. The approach to evaluate the framework included assessment against criteria for such dialogues and against some case studies concerning major political decisions.

An additional approach to evaluating an argumentation based decision support system is described in [73] where Sklar *et al.* propose an evaluation through a user study. The focus in [73] is to describe the user study conducted to evaluate the effectiveness of their proposed system: ArgTrust [61]. ArgTrust is a decision making tool based on a formal system of argumentation in which the evidence that influences a recommendation is modulated according to values of trust the user places on the evidence. In contrast to the situation this thesis focuses on, the aim was to assess the impact of ArgTrust on the users' decision making process, rather than ensuring that the tool and the user would obtain the same conclusion.

The framework proposed by Hunter *et al.* in [44] also includes a pilot study. The system proposed by Hunter *et al.* proposes a language to represent knowledge from clinical trials and incorporates clinical preferences over the types of evidence in order to aggregate the knowledge. The clinical pilot study described in [44] had two main aims: Firstly to validate the ability of their formalism to represent clinician preferences and secondly to see how stable the representation was. As part of the pilot study each clinician was presented with a document containing all the different evidence and types. They were then asked to decide which treatment they would recommend and give a reason for this. For each clinician and scenario an argument graph was generated and its extensions were compared to the actual treatment choices made by the clinician.

The end user evaluation is also key in the assessment of RecoMedic [57]. The aim of RecoMedic is to select the most appropriate items of medical literature given specific patient characteristics, this is achieved through the use of assumption based argumentation. The evaluation of this system was based on a set of medical studies relevant to the treatment of brain metastases, and was achieved though a user experience survey by a small group of postgraduate medical students.

In support of validating the contributions in this thesis one user study would be mainly used to confirm that the model recommendations made by the system match the ones made by a group of statisticians - all faced with the same data, research question and models in scope. This would bear some similarities to the user study in [44]. An additional user study similar to the user study approach proposed in [73], would be beneficial to assess end users' overall impression of the prototype, the ease of use and the user interface. This latter type of user study would be aimed mainly at clinicians (as the target end users) but would potentially benefit from being rolled out to statisticians too.

The evaluation methodologies used across these argumentation based decision support systems include the use of case studies, the development of a prototype, a specification and user studies. These methodologies are relevant to this thesis. Some are included as part of this thesis whilst others such as a user study necessitating prototype deployment and ethics approval are planned as part of future work.

7.2 Evaluation approach and timeline

A desirable set of criteria to evaluate the original contributions proposed in this thesis would include this following:

- Comprehensiveness and ability to deal with cases studies
- Implementability
- Scalability
- Maintainability
- Usability

The evaluation of the original contributions of this thesis can be achieved in different stages: evaluating the approach, evaluating the approach as implemented in a prototype through user studies and eventually as a full scale deployment on clinical data. The different stages and approaches to evaluation are illustrated in Figure 7.1, the shaded blocks represent the evaluation activities achieved within this thesis.

In an ideal scenario with no resource constraints a prototype system would be implemented on a set of clinical data sets of interest, and then a group of clinicians would be invited to use the system. A group of statisticians would also be required to validate that the recommendations and justifications generated by the prototype would align to their approach. This would require the availability of a working prototype, as well as time from clinicians and statisticians. In order to determine if the system is providing value by significantly reducing the time it takes to get from hypothesis to answer a comparison would need to be made either on the same research questions (one using this system and the other approaching it in other ways) or on the same clinicians feedback. This would benefit from a structure similar to one for a clinical trial in order

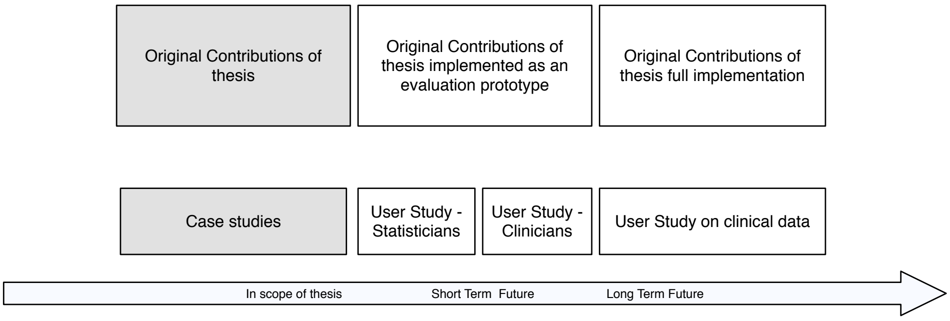


FIGURE 7.1: Evaluation Activities timeline

to isolate the benefit of the use of this system, from the differences in ways different users perceive the benefit of such an approach. A prototype would also enable all the evaluation criteria listed above to be assessed, assuming enough diverse data sets and users were available to the system.

A prototype has been developed as a web application by Zillesen [89] [1]. The feasibility of a roll out of this prototype to clinicians and statisticians within the time frame of this thesis is not realistic and as such different methods to evaluate this are proposed. The option of rolling it out to clinicians, pending identification of suitable clinicians, data sources and ethics approval remains a desirable option for future work. A roll out of a prototype in a clinical setting would assist in evaluating the usability of the proposed system from both functional and qualitative aspects.

The initial evaluation of the contributions proposed in this thesis is achieved through the use of case studies. These will assess whether the proposed approach is comprehensive enough to deal with different case studies and provide a similar model recommendation to the actual model implemented in each case study.

[1] Sign up and access on http://small-data-analyst.herokuapp.com

7.3 Empirical evaluation through case studies

7.3.1 Case Study 1: SENT

This data was collected as part of the SENT trial as introduced in section 1.3. The data available includes 415 patients with over 100 attributes (or variables) for each patient. There are a number of research questions that can be answered by leveraging this data.

In order to evaluate the case study the following research question (hypothesis) will be used:

Hypothesis 2

Is there a difference in overall survival between Female and Male patients?

The target variable is a column labelled OS which is of type "time to event". The contents of the SKB relevant to this scenario are in Appendix B. Therefore the initial step is to instantiate AS1 with objective $O = o_s$. This instantiates AS1 (Definition 4.1) which corresponds to the Z notation scheme $[FindModels]$.

The relevant contents of the SKB for this case study:

- $\{(m_{s1}, o_s), (m_{s2}, o_s), (m_{s3}, o_s), (m_{b1}, o_n), (m_{b2}, o_n),$
  $(m_{b3}, o_n), (m_{b4}, o_n), (m_{b5}, o_n), (m_{b6}, o_n)\} = F$
- $\{(m_{s1}, a_1), (m_{s2}, a_1), (m_{s3}, a_a), (m_{s2}, a_2),$
  $(m_{s3}, a_3), (m_{b1}, a_{10}), (m_{b2}, \neg a_{10}), (m_{b3}, a_8), (m_{b3}, a_{11}),$
  $(m_{b4}, a_9), (m_{b4}, a_{11}), (m_{b6}, a_8), (m_{b6}, a_{12})\} = C$
- $\{(o_s, o_n)\} = OBJ$

Arg_1 **Argument Scheme for model to consider on grounds of achieving the objective o_s : PM(o_s, r, m_{s1})**

Premise - o_s is the objective of the research question r

Premise - m_{s1} is able to analyse o_s

$\therefore$ - m_{s1} is suitable to answer r

Arg_2 **Argument Scheme for model to consider on grounds of achieving the objective o_s : PM(o_s, r, m_{s2})**

Premise - o_s is the objective of the research question r

Premise - m_{s2} is able to analyse o_s

$\therefore$ - m_{s2} is suitable to answer r

Arg_3 **Argument Scheme for model to consider on grounds of achieving the objective o_s : PM(o_s, r, m_{s3})**

Premise - o_s is the objective of the research question r

Premise - m_{s3} is able to analyse o_s

$\therefore$ - m_{s3} is suitable to answer r

The instantiation of AS1 has in this case study generated three arguments:

- Arg_1 : PM(o_s, r, m_{s1}): m_{s1}
- Arg_2: PM(o_s, r, m_{s2}): m_{s2}
- Arg_3: PM(o_s, r, m_{s3}): m_{s3}

The first critical question applies AS2 (Definition 4.2) in order to instantiate additional arguments in support of the use of models able to achieve the alternative objective to o_s. This invokes the [$FindAlternativeObjective$] scheme from Chapter 6.

Arg_4 **Argument for alternative objective: AO(o_s, r)**

- o_s is the objective of research question r
- o_n is an alternative objective to answer r
- m_{b1} calling $PM(o_n, r, m_{b1})$

$\therefore$ - m_{b1} is suitable to answer r

Arg_5 **Argument for alternative objective: AO(o_s, r)**

- o_s is the objective of research question r
- o_n is an alternative objective to answer r
- m_{b2} calling $PM(o_n, r, m_{b2})$

$\therefore$ - m_{b2} is suitable to answer r

Arg_6 **Argument for alternative objective: AO(o_s, r)**

- o_s is the objective of research question r
- o_n is an alternative objective to answer r
- m_{b3} calling $PM(o_n, r, m_{b3})$

$\therefore$ - m_{b3} is suitable to answer r

Arg_7 **Argument for alternative objective: AO(o_s, r)**

- o_s is the objective of research question r
- o_n is an alternative objective to answer r
- m_{b4} calling $PM(o_n, r, m_{b4})$

$\therefore$ - m_{b4} is suitable to answer r

Table 7.1: Results of assumptions tests for Case study 1

Assumption	Result
a_1	TRUE
a_2	TRUE
a_3	TRUE
a_8	TRUE
a_9	FALSE
a_{10}	TRUE
$\neg a_{10}$	FALSE
a_{11}	FALSE
$\neg a_{11}$	FALSE
a_{12}	FALSE

Arg_8 **Argument for alternative objective: AO(o_s, r)**

- o_s is the objective of research question r
- o_n is an alternative objective to answer r
- m_{b6} calling $PM(o_n, r, m_{b6})$

$\therefore$ - m_{b6} is suitable to answer r

This results in six additional arguments in support of the use of additional models. For each of these arguments in the current set of arguments:

$$\{Arg_1, Arg_2, Arg_3, Arg_4, Arg_5, Arg_6, Arg_7, Arg_8\}$$

The second critical question ($CQ2$) tests the critical assumptions thereby undercutting the argument in support of the use of a model when the critical assumptions for that model don't hold.

The assumptions to be tested are: $\{a_1, a_2, a_3, a_8, a_9, a_{10}, \neg a_{10}, a_{11}.\neg a_{11}, a_{12}\}$. The mapping of assumptions to their respective models is detailed in 7.3.1. The result of the assumptions testing is in table 7.1.

The instantiation of the AS3 (Definition 4.3) Argument Scheme $CA(m_i)$ results in the following:

Arg_9 **Argument against the use of a Model for failed critical assumption: CA(**(m_{b2})

- Model m_{b2} achieves objective o_n
- $\neg a_{10}$ is a critical assumption for m_{b2}
- $\neg a_{10}$ does not hold

$\therefore$- m_{b2} is not a model to be considered

An additional three arguments are instantiated by $CA(m_i)$ as above: $Arg_{10} : \neg m_{b3}, Arg_{11} : \neg m_{b4}, Arg_{12} : \neg m_{b6}$ resulting in the following set of arguments:

- Arg_1 : PM(o_s, r, m_{s1}): m_{s1}
- Arg_2: PM(o_s, r, m_{s2}): m_{s2}
- Arg_3: PM(o_s, r, m_{s3}): m_{s3}
- Arg_4: AO(o_s, r): m_{b1}
- Arg_5: AO(o_s, r): m_{b2}
- Arg_6': AO(o_s, r): m_{b3}
- Arg_7: AO(o_s, r): m_{b4}
- Arg_8: AO(o_s, r): m_{b6}
- Arg_9: CA$(((m_{b2})$: $\neg m_{b2}$
- Arg_{10}: CA$(((m_{b3})$: $\neg m_{b3}$
- Arg_{11}: CA$(((m_{b4})$: $\neg m_{b4}$

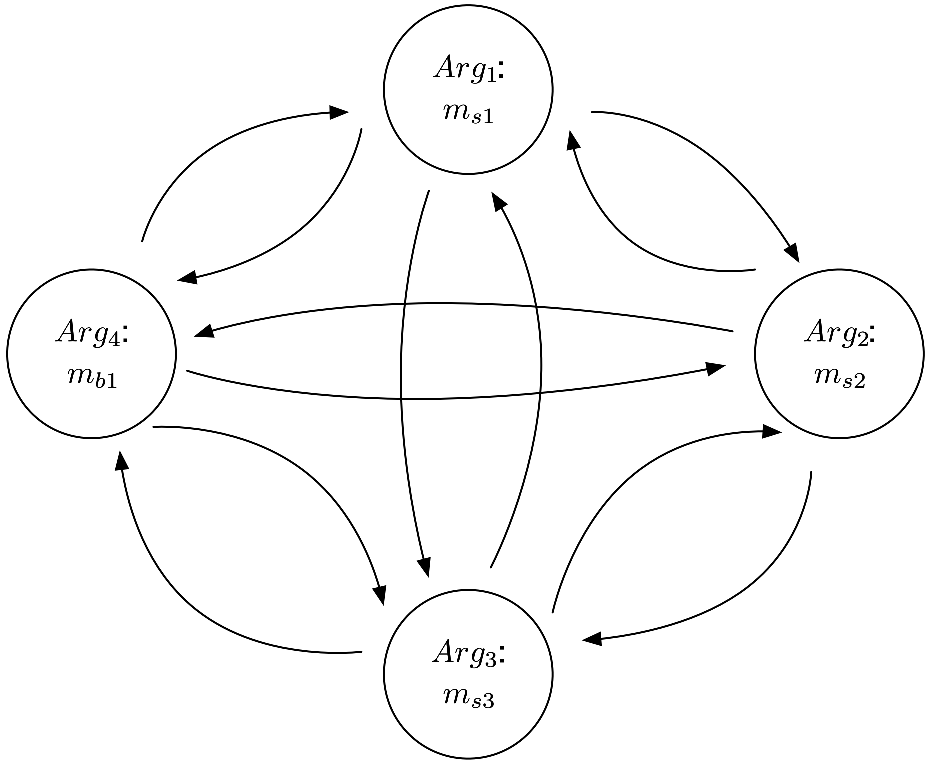


FIGURE 7.2: Argumentation Framework for Case Study 1

- Arg_{12}: CA(((m_{b6}) : $\neg m_{b6}$

The arguments in favour of the use of a model that are not attacked within the argumentation framework are:

- Arg_1 : PM(o_s, r, m_{s1}): m_{s1} (m_{s1} is *Kaplan-meier*)
- Arg_2: PM(o_s, r, m_{s2}): m_{s2} (m_{s2} is *Cox Proportional Hazards*)
- Arg_3: PM(o_s, r, m_{s3}): m_{s3} (m_{s3} is *Weibull*)
- Arg_4: AO(o_s, r): m_{b1} (m_{b1} is χ^2)

context domain	model m	Performance measure p
$cd1_3$ heavy censoring	m_{s1} KM	p_3 = strongly affected
	m_{s2} PH	p_1 = unaffected
	m_{s3} $Weibull$	p_3 =strongly affected
	m_{b1} χ^2	p_3 =strongly affected

TABLE 7.2: $cd1_3$ performance function mapping relevant to Case Study 1: SENT

Assuming the objective is to run the most suitable models then the preferences derived from the relevant context domains are to be generated. There are two relevant contexts in this case: $cd1_3$ is heavy censoring and $cd2_2$ is model intent explain. $cd1_3$ is relevant as in this data 80 % of the cases are censored.

Referring to Table 7.2 the preference order over models resulting from this is $ph \succ \{km, wei.\chi^2\}$. Using Definition 5.2 the three meta-level preference arguments are generated to reflect this:

- $(PA_{cd1-12}, (m_{s1}, m_{s2}))$ as $p_{cd1}(m_{s1}) \prec p_{cd1}(m_{s2})$
- $(PA_{cd1-13}, (m_{s3}, m_{s2}))$ as $p_{cd1}(m_{s3}) \prec p_{cd1}(m_{s2})$
- $(PA_{cd1-23}, (m_{b1}, m_{s2}))$ as $p_{cd1}(m_{b1}) \prec p_{cd1}(m_{s2})$
- where $PA_{cd1-12}, PA_{cd1-13}, PA_{cd1-23} \in \mathcal{D}$

These meta-level arguments are now added to the argumentation framework in Figure 7.2 to produce the extended argumentation framework in Figure 7.3.

The preference level arguments attack the attack of Arg_1, Arg_3, Arg_4 on Arg_2. Therefore Arg_2 is no longer attacked by the other arguments. This results in $Arg_2 : [ph]$ being acceptable with respect to the preference arguments from $cd1_3$ for this EAF. Therefore the recommended model to apply given heavy censoring is ph.

Including also the second relevant context domain ($cd2_2$), by using the mapping of performance measures to models in Table 7.3 generates a different preference ordering

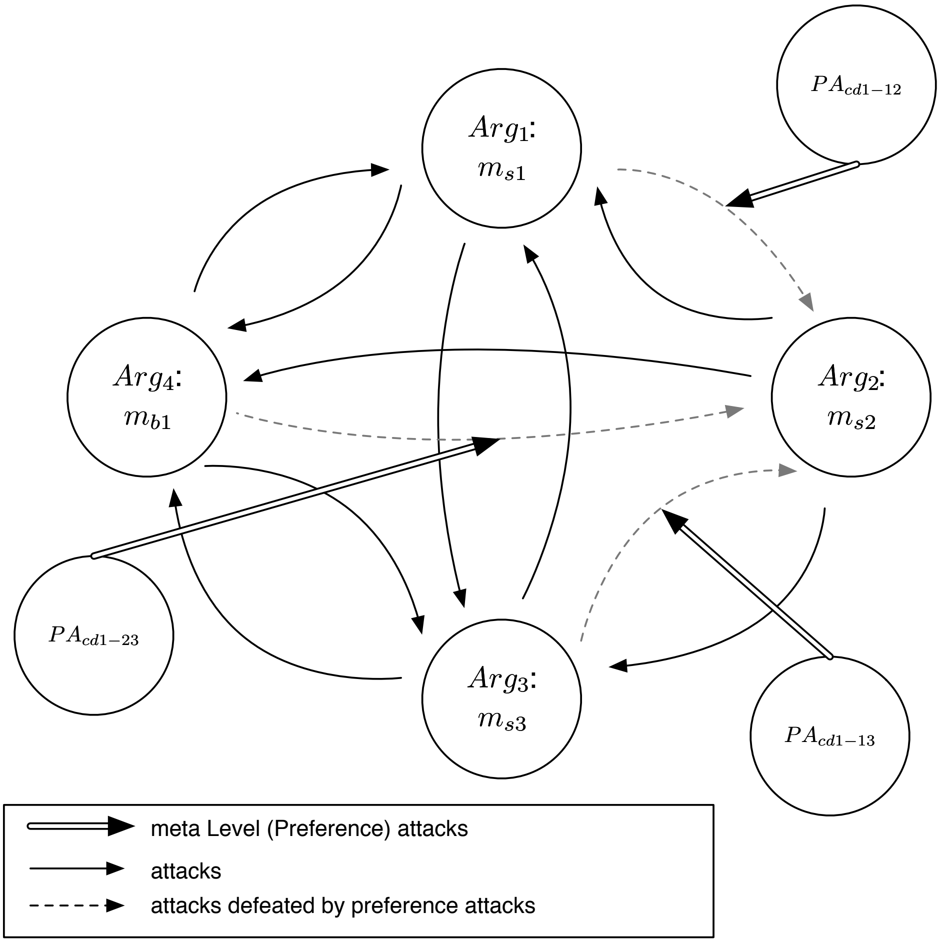


FIGURE 7.3: Extended Argumentation Framework for Case Study 1

context domain	model m	Performance measure p
$cd2_2$ explain	m_{s1} KM	p_1 suitable
	m_{s2} PH	p_1 suitable
	m_{s3} Weibull	p_2 neutral
	m_{b1} χ^2	p_2 neutral

TABLE 7.3: $cd2_2$ performance function mapping for model intent ='explain' relevant to Case Study 1: SENT

on the models. Given the intent of the analysis is to explore the data (rather than predict) then $m_{s1}(km), m_{s2}(ph)$ are both preferred to achieve this intent over the other models considered. The four meta level preference arguments generated are:

- $(PA_{cd2-12}, (m_{s3}, m_{s2}))$ as $p_{cd2}(m_{s3}) \prec p_{cd2}(m_{s2})$
- $(PA_{cd2-13}, (m_{s3}, m_{s1}))$ as $p_{cd2}(m_{s3}) \prec p_{cd2}(m_{s1})$
- $(PA_{cd2-22}, (m_{b1}, m_{s2}))$ as $p_{cd2}(m_{b1}) \prec p_{cd2}(m_{s2})$
- $(PA_{cd2-23}, (m_{b1}, m_{s1}))$ as $p_{cd2}(m_{b1}) \prec p_{cd2}(m_{s1})$
- where $PA_{cd2-12}, PA_{cd2-13}, PA_{cd2-22}, PA_{cd2-23} \in \mathcal{D}$

Figure 7.4 illustrates the argumentation framework resulting from meta level arguments from contexts $cd1_3$ and $cd2_2$. The preferred extension for this argumentation framework includes Arg_2 and the meta-level arguments derived from contexts $cd1_3$ and $cd2_2$. This results in the recommended model $m_{s2} : ph$ on the grounds that it is supported by $Arg_2 : m_{b2}$ and acceptable to the EAF that includes the meta level arguments generated from $cd1_3$ and $cd2_2$. The analysis approach used to answer Hypothesis 2 in the case study was proportional hazards m_{s2} which is consistent with the recommendation obtained by applying the argument schemes, critical questions and extended statistical knowledge base as proposed in this thesis.

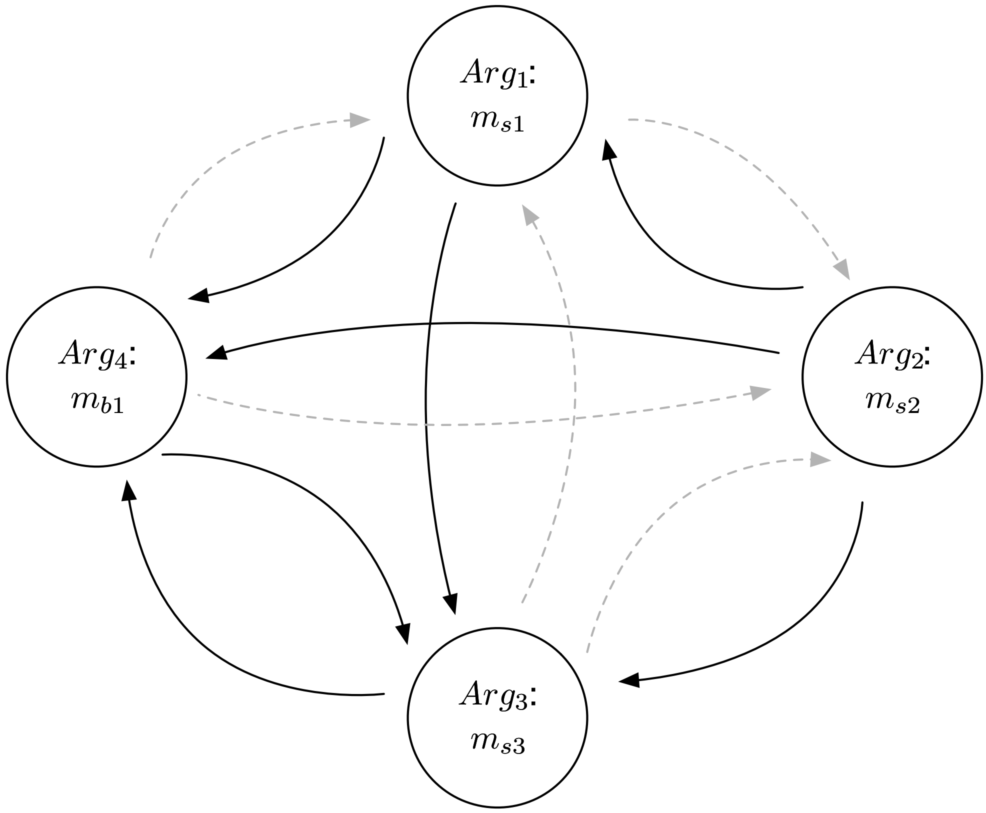


FIGURE 7.4: Argumentation Framework for Case Study 1 resulting from meta level arguments from $cd1_3$ and $cd2_2$

7.3.2 Case Study 2: Benchmarking Complications

This data set consists of a larger cohort of patients and is a result of a retrospective exercise in data collection. The relevant clinical records concerning 1034 SCC (Squamous Cell Carcinoma) surgeries from three different hospitals have been collected. The analysis cohort included only patients' first surgery and removed any cases with missing outcome data. This resulted in an analysis data set comprising 802 patients' first surgery. The analysis objective or hypothesis was to predict the likelihood of a patient experiencing a post operative complication following SCC surgery. A post operative complication is defined as a complication occurring within 30 days of surgery.

Hypothesis 3

Predict the likelihood of a patient experiencing a post operative complication following SCC surgery

The available data included 40 attributes that were deemed to be clinically relevant and after an initial review of these the attributes of potential interest were reduced to 20. These are listed in Table 7.4.

In this case study the target variable of interest (comp_30) was binary, the relevant objective for a binary target variable is o_n. The number of covariates available was large. The application of the proposed methodology in this setting would assist in recommending the most suitable approach. Note that in this example there is no relevant alternative objective.

The SKB (Definitions 3.1 and 3.2) contents of relevance to this case study are:

- $\{(m_{b1}, o_n), (m_{b2}, o_n), (m_{b3}, o_n), (m_{b4}, o_n), (m_{b5}, o_n),$
  $(m_{b6}, o_n)\} = F$

TABLE 7.4: Complications Data Set

Name	Type	Description
comp_30	Binary	Indicator of complications
group	Nominal	Hospital
age	Interval	Age of patient at surgery
sex	Binary	Gender of patient
alcohol	Nominal	Patients' drinking habits
smoking	Nominal	Patient's smoking habits
cvs	Binary	Patient's cardiovascular comorbidity
resp	Binary	Patient's respiratory comorbidity
abdo	Binary	Patient's gastro intestinal comorbidity
performance	Nominal	Patient's physical status
prevt	Binary	Indicator of previous treatment
site	Nominal	Site of tumour
t	Nominal	Grading of tumour
n	Nominal	Grading of nodes
margins	Nominal	Margins of tumour
ecs	Binary	Extra Capsular Spread
scale	Nominal	Scale of Surgery
flap	Binary	Flap indicator
trach	Binary	Tracheostomy
a.time.hours	Interval	anaesthetic time
bloodloss	Interval	Blood loss from surgery in litres

- $\{(m_{b1}, a_{10}), (m_{b1}, \neg a_{11}), (m_{b2}, \neg a_{10}), (m_{b3}, a_8), (m_{b3}, a_{11}),$
  $(m_{b4}, a_9), (m_{b4}, a_{11}), (m_{b6}, a_8), (m_{b6}, a_{12})\} = C$

The instantiation of the AS1(Definition 4.1) given the objective of o_n would be as follows:

Arg_1 **Argument Scheme for model to consider on grounds of achieving the objective m_{b1} : PM(o_n, r, m_{b1})**

Premise -	o_n is the objective of the research question r
Premise -	m_{b1} is able to analyse o_n
$\therefore$ -	m_{b1} is suitable to answer r

The instantiation of AS1 results in the following set of arguments:

- Arg_1 : PM(o_n, r, m_{b1}): m_{b1}
- Arg_2: PM(o_n, r, m_{b2}): m_{b2}
- Arg_3: PM(o_n, r, m_{b3}): m_{b3}
- Arg_4 : PM(o_n, r, m_{b4}): m_{b4}
- Arg_5: PM(o_n, r, m_{b5}): m_{b5}
- Arg_6: PM(o_n, r, m_{b6}): m_{b6}

No alternative objectives are defined for o_n within the SKB therefore the instantiation of CQ1(Definition 4.2) does not result in any additional arguments.

The instantiation of CQ2 (Definition 4.3) validates the assumptions relevant to each of the models supported by arguments. The results of the assumptions testing on this data and situation is available in Table 7.5.

Table 7.5: Assumptions for Case study 2

Assumptions	Result
a_8	TRUE
a_9	TRUE
a_{10}	TRUE
$\neg a_{10}$	FALSE
a_{11}	TRUE
$\neg a_{11}$	FALSE
a_{12}	FALSE

Instantiating CQ2

- Model m_{b1} achieves objective o_n
- $\neg a_{11}$ is a critical assumption for m_{b1}
- $\neg a_{11}$ does not hold

$\therefore$- m_{b1} is not a model to be considered

The instantiation of CQ2 (Definition 4.3) generates arguments against the use of some models: $\{Arg_7 : \neg\, m_{b1}, Arg_8 : \neg\, m_{b2}, Arg_9 : \neg\, m_{b6}\}$. As a result of these there will be only three arguments that are not been undercut by the critical question CQ2: $\{Arg_3 : m_{b3}, Arg_4 : m_{b4}, Arg_5 : m_{b5}\}$. These arguments and their attack relationships can be seen in the argumentation framework in Figure 7.5. Under the assumption that only the most suitable models should be used then the context domains can be used to recommend the most suitable models.

In this situation the declared intent of the analysis was to predict ($cd2_1$). The mapping of the performance measures and models relevant to this case study is in Table 7.6.

Applying Definition 5.2 this results in the following meta-level preference arguments:

- $(PA_{cd2-13}, (m_{b4}, m_{b3}))$ as $p_{cd2}(m_{b4}) \prec p_{cd2}(m_{b3})$

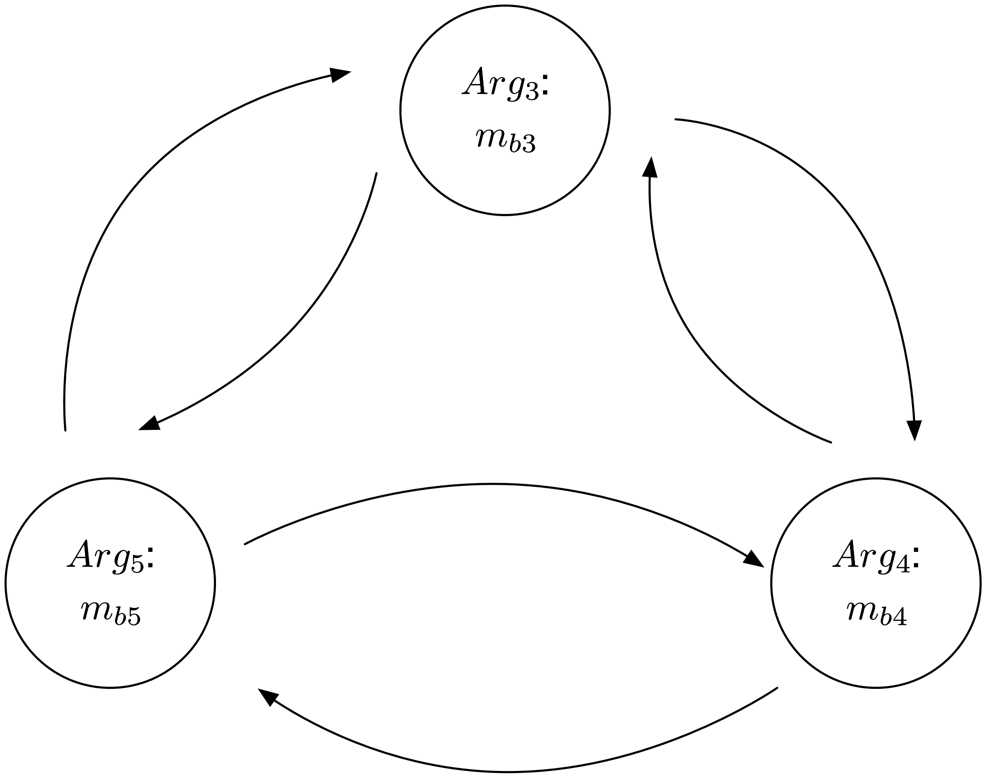


Figure 7.5: Argumentation Framework for Case Study 2

context domain	model m	Performance measure p
$cd2_1$ predict	m_{b3} LR	p_1 suitable
	m_{b4} DT	p_2 neutral
	m_{b5} NN	p_1 suitable

Table 7.6: $cd2_1$ performance function mapping for model intent ='predict' relevant to Case Study 2

- $(PA_{cd2-12}, (m_{b4}, m_{b5}))$ as $p_{cd2}(m_{b4}) \prec p_{cd2}(m_{b5})$
- where $PA_{cd2-13}, PA_{cd2-12} \in \mathcal{D}$

The effect of applying the preference arguments $\{PA_{cd2-13}, PA_{cd2-12}\}$ from the context domain $cd2_1$ to the argumentation framework generates the EAF in Figure 7.6. This does not result in an admissible extension, other than the empty set, as $Arg_3 : m_{b3}$ and $Arg_5 : m_{b5}$ symmetrically attack each other.

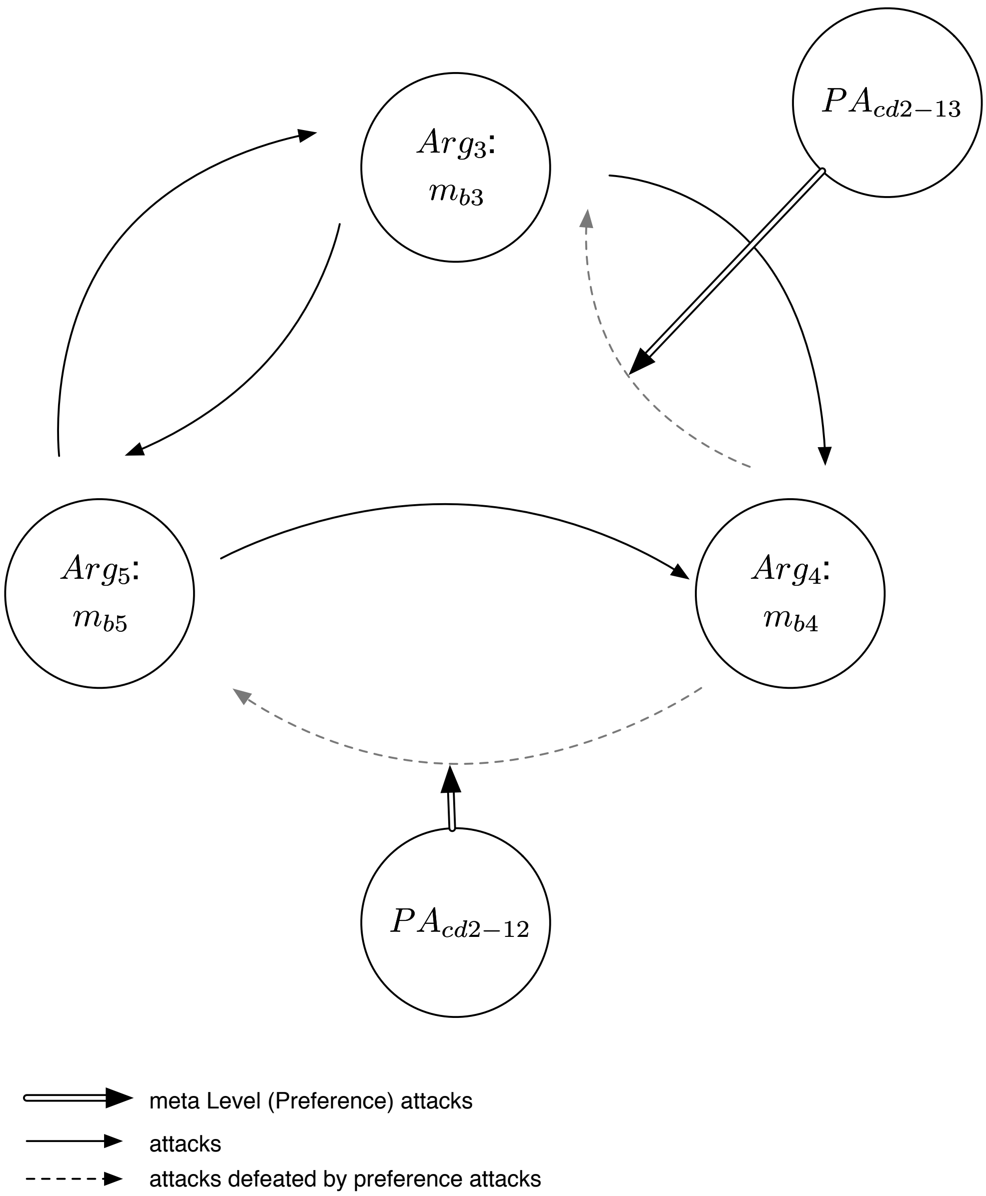


FIGURE 7.6: Extended Argumentation Framework for Case Study 2 with $cd2_1$

context domain	model m	Performance measure p
$cd3$ missing data	m_{b3} LR	p_2 unsuitable
	m_{b4} DT	p_1 suitable
	m_{b5} NN	p_2 unsuitable

TABLE 7.7: $cd3$ performance function mapping for missing data relevant to Case Study 2

The second relevant context domain ($cd3$) is missing data. The mapping of the performance measures and models relevant to this case study is in Table 7.7.

Note that the definition employed in this case study for evaluating the extent of missing data is simplistic and solely base on the substantial percentage (close to 80 %) of non complete cases (i.e. cases with at least one item of data is missing). The preference orders over the models from Table 7.7 result in the following meta-level preference arguments:

- $(PA_{cd3-13}, (m_{b3}, m_{b4}))$ as $p_{cd3}(m_{b3}) \prec p_{cd3}(m_{b4})$
- $(PA_{cd3-12}, (m_{b5}, m_{b4}))$ as $p_{cd3}(m_{b5}) \prec p_{cd3}(m_{b4})$
- where $PA_{cd3-13}, PA_{cd3-12} \in \mathcal{D}$

Applying only the preference arguments generated from $cd3$ results in an EAF as illustrated in Figure 7.7. The results in $\{Arg_4 : m_{b4}\}$ being acceptable with respect to the set of preference arguments resulting from $cd3$.

The preference arguments from $cd2_1$ are in direct contradiction with those resulting from $cd3$. In this particular situation $cd2_1 \succ cd3$ and the decision to explore the use of logistic regression (m_{b3}) was the first approach applied.

In this case study there was a clinician expressed preference for the use of m_{b3}. The following meta level arguments can be generated from this expressed preference ($cd4$) using Definition 5.2:

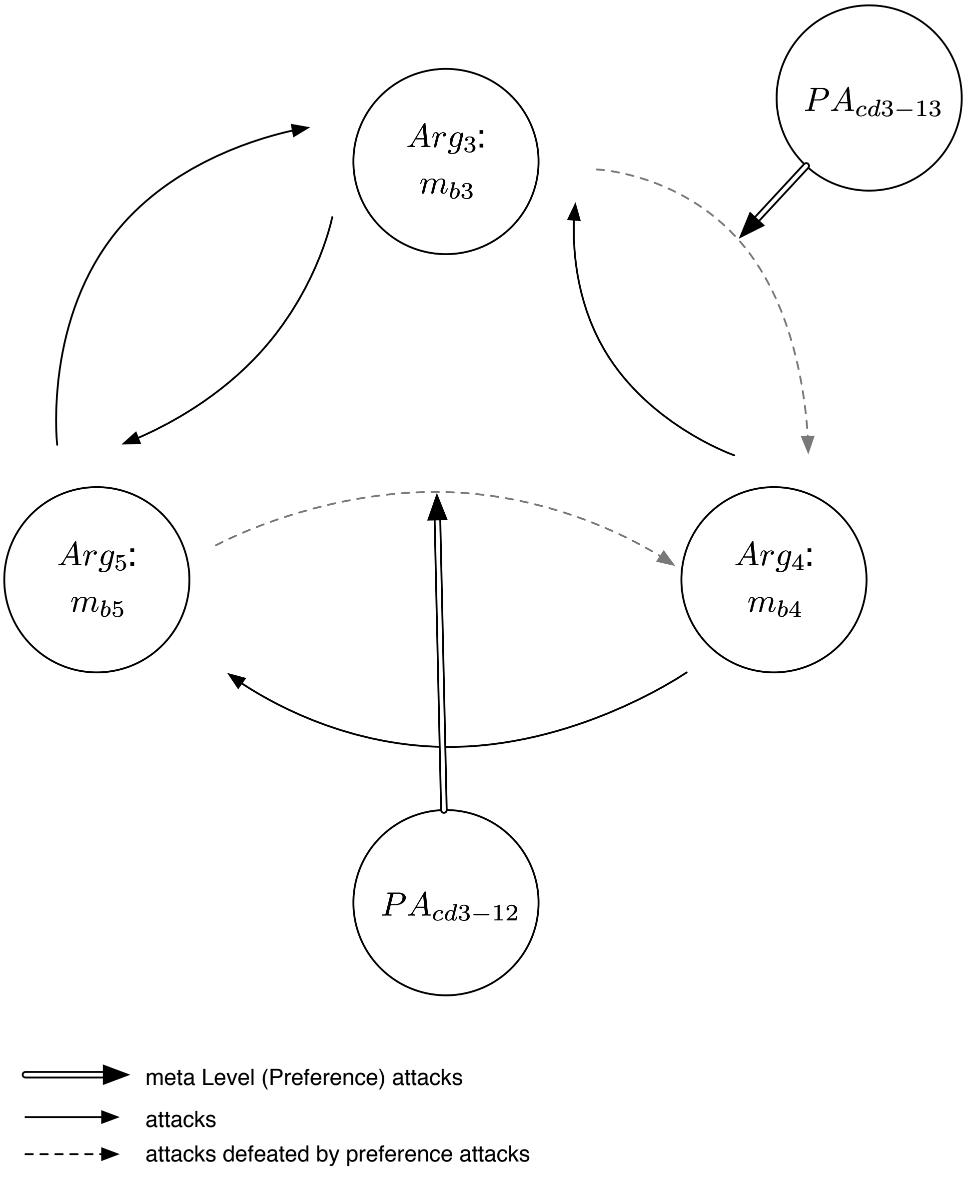


FIGURE 7.7: Extended Argumentation Framework for Case Study 2 with $cd2$

- $(PA_{cd4-13}, (m_{b4}, m_{b3}))$ as $p_{cd4}(m_{b4}) \prec p_{cd4}(m_{b3})$
- $(PA_{cd4-12}, (m_{b5}, m_{b3}))$ as $p_{cd4}(m_{b5}) \prec p_{cd4}(m_{b3})$
- where $PA_{cd4-13}, PA_{cd4-12} \in \mathcal{D}$

The introduction of clinician preference $cd4$, which in this case study was a preference for the use of m_{b5}(LR), was also considered. Given the meta-level arguments generated from $cd4$ the preferred extension with respect to the meta level arguments generated from $cd2_1$ and $cd4$ contained the meta-level arguments from $cd2_1$ and $cd4$ and $\{m_{b3}\}$. In effect the justification for the use of m_{b3} over the other possible models was based on the argument in support of its use to achieve the objective, a preference generated from the intent of the analysis to predict and clinician preference for the use of m_{b3} which is *Logistic Regression* (LR). This is illustrated in Figure 7.8.

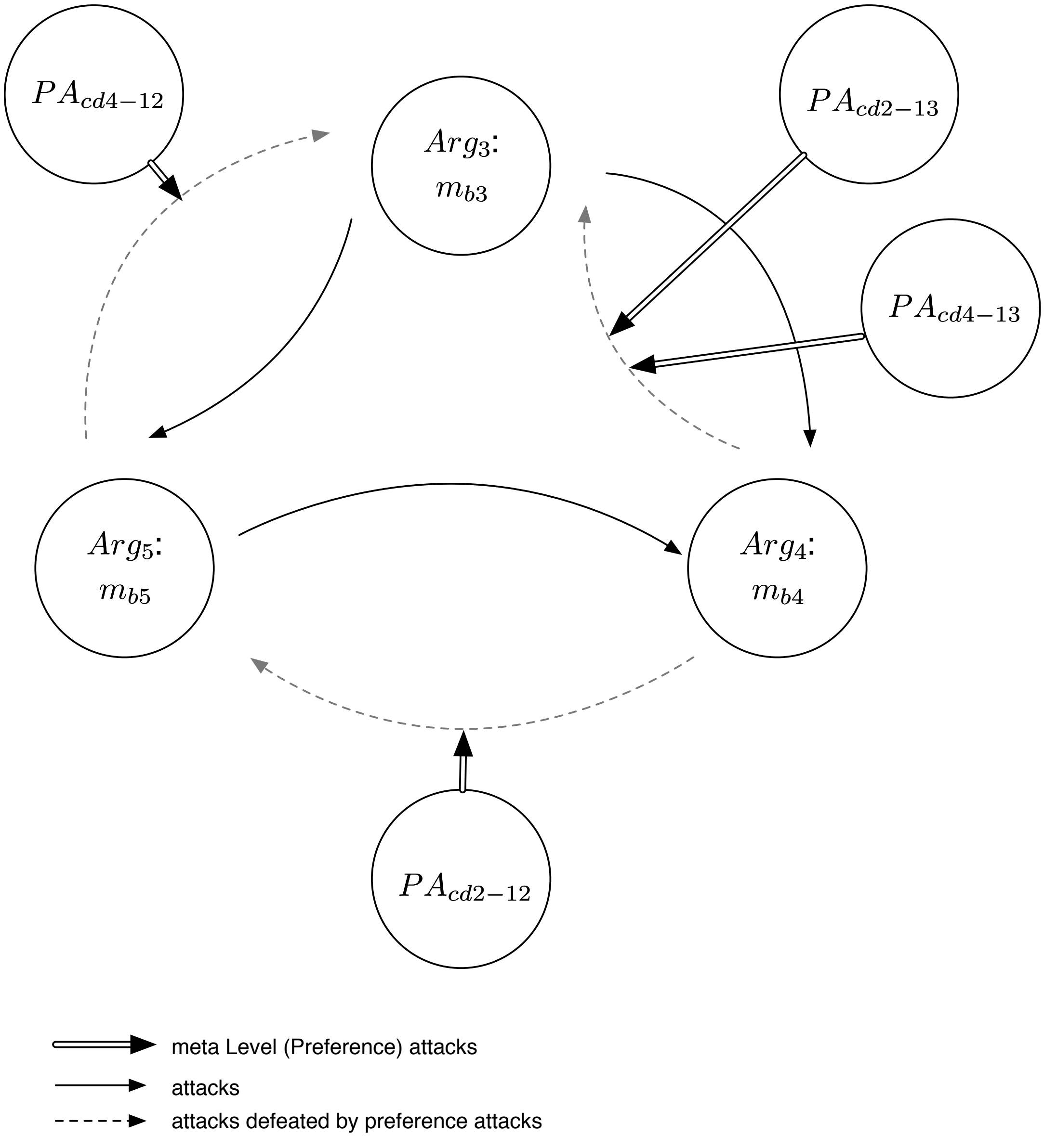


FIGURE 7.8: Extended Argumentation Framework for Case Study 2 with $cd2_1$ and $cd4$

7.3.3 Conclusion from the Case Studies

The worked out case studies indicate that the breadth of considerations made related to selecting the most appropriate model to apply in order to answer a research question were able to be accommodated when applying the argument schemes, critical questions and extended SKB as proposed in this thesis. The recommended models were in line with the models applied in the case studies.

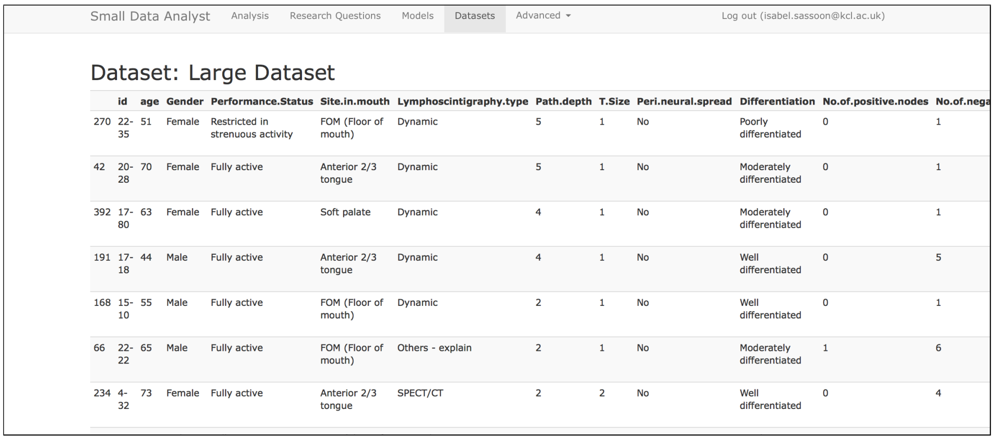
Although the case studies are indicative of the type of scenarios where such an approach would be applied, more evaluation as illustrated in Figure 7.1 is required. The case studies provide an initial positive step towards more comprehensive evaluation.

7.4 Prototype design and evaluation

A prototype of the methodology was implemented by [89] as part of the requirements for the MSc Web Intelligence at King's College London. The prototype was designed to implement the methodologies proposed in this thesis [2].

The SENT data as introduced in Section 1.3 in a modified and anonymised format was made available to the MSc project in order for it to be used as an example when demonstrating the functionality and design of the developed app. The main difference between the developed app and the empirical evaluation through case studies presented in Section 7.3.1 is that the app did not offer an alternative objective and some restrictions on the type of data that can be uploaded onto the hosted platform where the prototype resides - therefore only an anonymised version of the SENT data was included.

[2] Sign up and access on `http://small-data-analyst.herokuapp.com`


Small Data Analyst Analysis Research Questions Models Datasets Advanced Log out (isabel.sassoon@kcl.ac.uk)

Dataset: Large Dataset

	id	age	Gender	Performance.Status	Site.in.mouth	Lymphoscintigraphy.type	Path.depth	T.Size	Peri.neural.spread	Differentiation	No.of.positive.nodes	No.of.nega
270	22-35	51	Female	Restricted in strenuous activity	FOM (Floor of mouth)	Dynamic	5	1	No	Poorly differentiated	0	1
42	20-28	70	Female	Fully active	Anterior 2/3 tongue	Dynamic	5	1	No	Moderately differentiated	0	1
392	17-80	63	Female	Fully active	Soft palate	Dynamic	4	1	No	Moderately differentiated	0	1
191	17-18	44	Male	Fully active	Anterior 2/3 tongue	Dynamic	4	1	No	Well differentiated	0	5
168	15-10	55	Male	Fully active	FOM (Floor of mouth)	Dynamic	2	1	No	Well differentiated	0	1
66	22-22	65	Male	Fully active	FOM (Floor of mouth)	Others - explain	2	1	No	Moderately differentiated	1	6
234	4-32	73	Female	Fully active	Anterior 2/3 tongue	SPECT/CT	2	2	No	Well differentiated	0	4

FIGURE 7.9: Data in the prototype

The availability of a prototype (with access to real data and ethics approval) would enable the usability to be evaluated through user studies and additional validation of the underlying knowledge base, schemes and inference by comparing the prototype recommendations with a statistician's one. The implementation of the prototype in a clinical setting would also evaluate the implementation process, the maintainability of the system and its scalability.

Case Study 1: SENT in the prototype

The data used in the prototype is an anonymised version of the SENT trial data introduced in Section 1.3 and matches the analysis empirically evaluated in Section 7.3.1. In order for this to be used in the prototype all demographic and personally identifiable columns were removed (such as date of birth). Furthermore the most clinically relevant columns were selected. This resulted in a smaller sized data table including 25 columns. The data is available to visually inspect in the Dataset tab, see Figure 7.9.

FIGURE 7.10: Starting a new analysis in the prototype

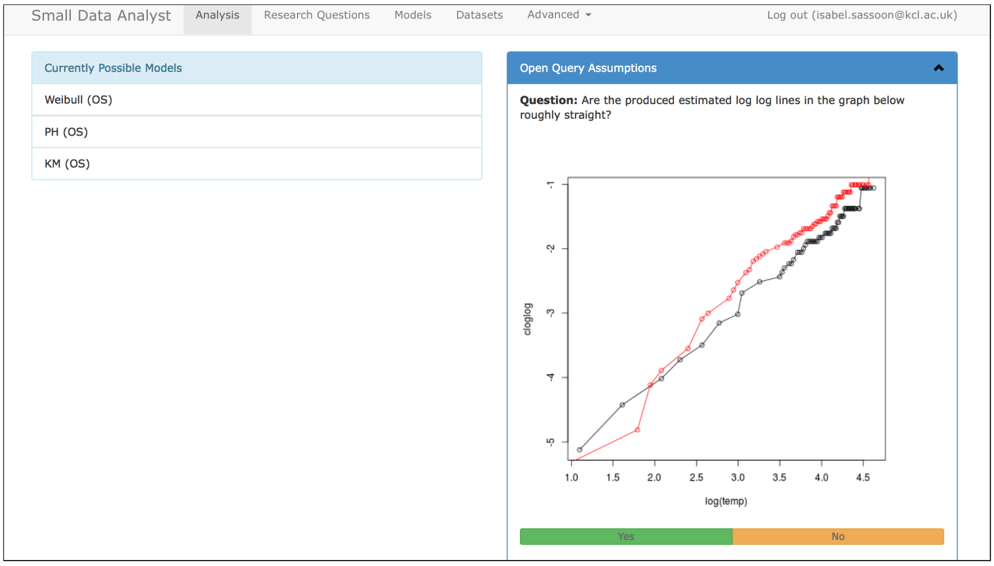


FIGURE 7.11: Clinician answering the first question on the assumptions in the prototype

In order to start an analysis to answer the research question (Hypothesis 2): Is there a difference in the survival (OS) by gender? then Figure 7.10 illustrates the steps are taken in the GUI of the prototype.

Figure 7.11 and 7.12 illustrate how the answer to the critical questions is elicited from the clinician.

As the first assumption question is answered by the clinician by selecting either 'yes' or 'no', the next query is displayed under it and the list of possible models is updated

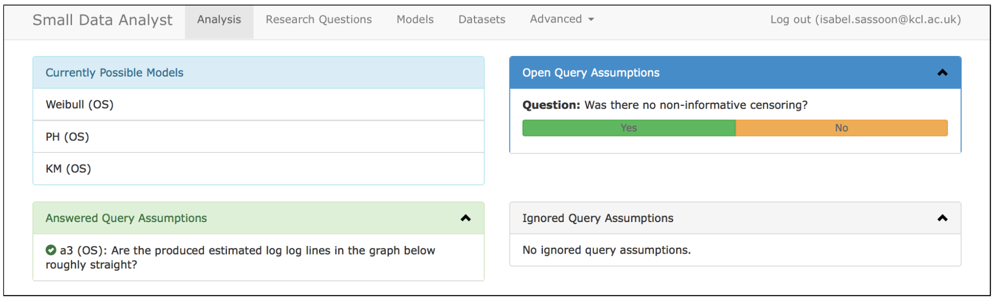


FIGURE 7.12: Clinician answering a subsequent question on the assumptions in the prototype

if required. In this case the answer of 'yes' to the critical question does not generate any arguments against the use of any model therefore the list is unchanged. The next question to be answered by the clinician is in Figure 7.12.

Following the confirmation by the clinician that the weibull assumption held and that there was no non-informative censoring the prototype recommends the use of the model *ph*, and the reasons for this are listed under answered query assumptions. This is in Figure 7.13.

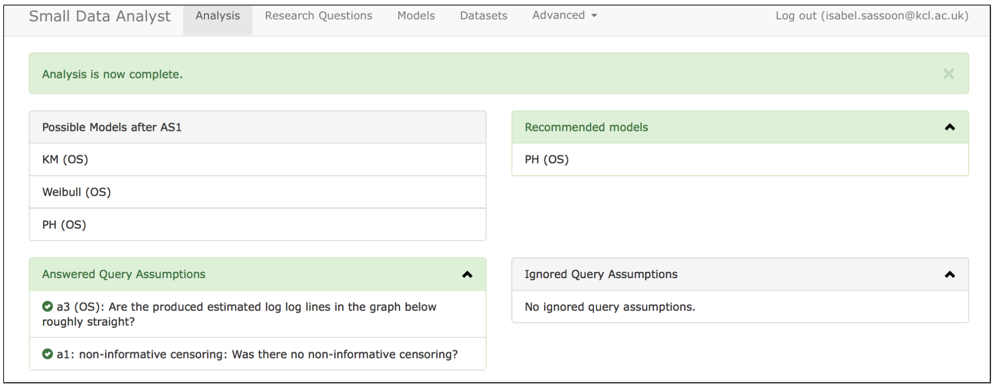


FIGURE 7.13: Displaying the recommended model in the prototype

Within the prototype there is also the option to visualise the extended argumentation framework used to generate the recommendation, see Figure 7.14. This recommendation is based on the context domain related to censoring as defined in the statistical knowledge base (the context domain is defined by the admin user and is based on the contents of the SKB as in Appendix B). The relevance of this context domain is based on the proportion of censored cases within the dataset, and this can be automatically generated from the dataset.

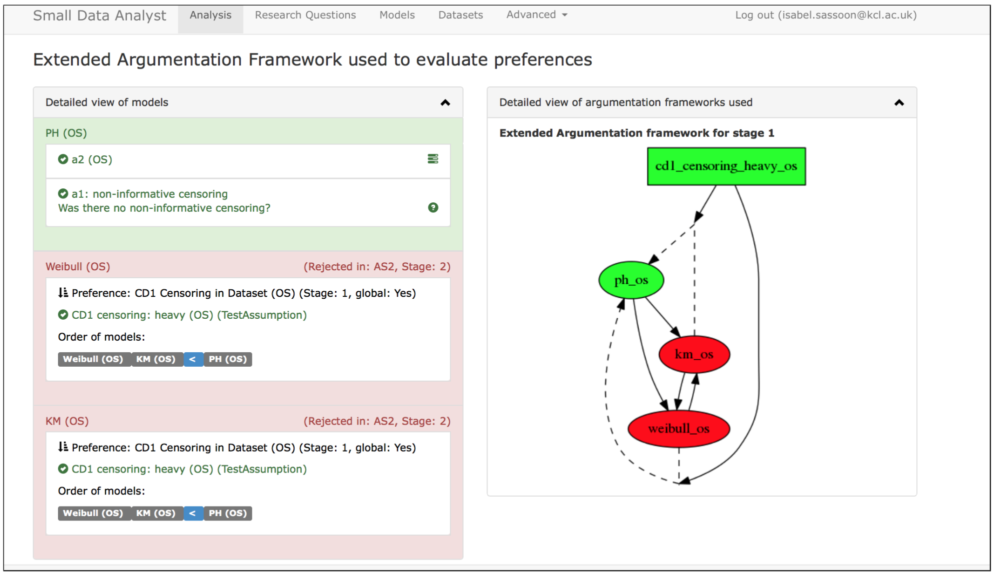


FIGURE 7.14: Displaying the recommended model in the prototype

7.5 Summary

The criteria to evaluate the methodology proposed in this thesis have been initially assessed through the use of case studies. The comparison of the results of the empirical case studies compared to the results obtained applying the proposed approach have shown that, for the defined knowledge base, the system is able to replicate the process and match the outcome. The development of a prototype based on the Z notation specifications in Chapter 6 is an initial step in assessing the feasibility of the implementation of the specifications based on the original contributions articulated in this thesis.

In the absence of a roll out of the prototype and in order to evaluate the comprehensiveness of the proposed method an external validation of the recommendations of the models would be preferred. This would assist in validating not only that the prototype system based on the original contribution of this thesis produces a recommendation but that the resulting recommendation is in line with the reasoning a statistician would employ given the same research question, data and circumstances.

As discussed in section 7.1 and 7.2 there are additional aspects of the evaluation process that have yet to be undertaken. The next step of the evaluation would be the roll out of the prototype on indicative clinical data in order to complete two user studies. One would be addressed to statisticians and would involve documenting the model recommendations a statistician would make, given the same research questions and data sets in the prototype, and validating that the reasoning is consistent. The second user study would be directed at clinicians and would look at collecting feedback on the system as well as measuring the benefit of the system in terms of *time to result* compared to previous working patterns.

In order to test the scalability and maintainability of the proposed approach the structure of the underlying knowledge base needs be expandable, modifiable as I anticipate that this is an area where the application of the approach proposed in this thesis on a different domain will require this. This can be achieved through the design principle of the specification proposed in Chapter 6 that includes schemes to maintain the SKB as well as alerts in case of errors.

Chapter 8

Conclusions and Future Directions

In this chapter I summarise the approach proposed for statistical model selection emphasising the contributions made in this thesis.

In this thesis I have developed and articulated an approach that provides recommended statistical model(s) given a clinician's research question (hypothesis) and data. The similarity in the process of selecting the most appropriate statistical model to use to that of diagnosing a patient, and the challenges related to conflicting conclusions, incomplete information and the necessity to justify any recommendations have resulted in an argumentation based approach.

The proposed approach relies on the following original contributions: extended statistical knowledge base, the argument scheme and critical questions to generate an argumentation framework which enables the recommendation and justification in support of the use of the most suitable model(s). The methodology also allows for the inclusion of preferences, deriving from different contexts and their inclusion in the argument framework evaluation through the use of Extended Argumentation Frameworks.

The methodology allows for inputs and information to be gleaned from the clinician (end user) as well as from the data. Preferences derived from statistical theory, model intent and personal preferences of the clinicians (end user) are also leveraged as part of the recommendation process. The justification for the recommendation of a model (or set of models) over others is delivered as part of the argument structure.

The objective I set out to address was to devise a method to recommend the most appropriate statistical model(s) given a clinician's research question and data. The method proposed, its formalisation and evaluation through case studies illustrate that the system can provide a model recommendation. The translation of the Z notation specification to a prototype illustrates the feasibility of implementation and provides a basis for future evaluation by end users and statisticians as suggested in Chapter 7.

In Section 8.1 I review the original contributions made in this thesis. In Section 8.2 I focus on research challenges that are related to the topic explored in this thesis that I plan to focus on in the future.

8.1 Summary of the Contributions

This thesis makes for the following contributions:

- **The role of argumentation in diagnosing the appropriate statistical model to be used given a research question and the available data**. The process of recommending the most appropriate statistical model(s) was formulated in a format suitable to be supported through arguments and argumentation.

This formed the foundation of the methodology proposed in this thesis that recommends the most appropriate statistical model to apply given the clinician's research question and the available data.

- **The structure of the argumentation scheme and its associated critical questions** The process of ascertaining the suitability of a model was implemented as an argumentation scheme that leveraged both a knowledge base and the clinician as it is instantiated. This also achieved the separation of the process of generating arguments in support of the use of a model from the knowledge base. This enables changes and expansions to be made in the knowledge base without affecting the scheme. The relevant critical questions were articulated and are themselves represented as argumentation schemes.

- **The structure and role of the Statistical Knowledge Base** The elements required to support the choice of model were selected. The structure of the knowledge base was designed to support the instantiation of the argument scheme and critical questions.

- **The role and source of preferences in the context of statistical model selection** The different sources of preference orders that are relevant to the process of statistical model selection were categorised into three groups: feature based, intent based and domain based. In order to differentiate between the different preference orders over models a notion of context domain was defined.

- **The structure of the extended knowledge base** The extension of the knowledge base was devised to include context domains, performance measures and mappings in support of the preference orders. Each context domain has an associated set of performance measures and a mapping between these and models.

- **Reasoning with Arguments and context domain derived preference arguments through Extended Argumentation Frameworks** The use of the extended knowledge base was coupled with Extended Argumentation Frameworks to enable reasoning with arguments and preference arguments, applying the relevant context domain for preference importance order. This enables the recommendation of the most suitable model or models, when more than one model is possible and the set of context domains applied provide the justification for the recommendation.

- **Formalisation of the system** The proposed original contributions were formalised in Z notation. The Z notation specifications were leveraged in the development of a prototype.

- **Evaluation** The criteria for the evaluation of similar systems were reviewed and those relevant to this proposed methodology were considered. The initial evaluation of the contributions proposed within this thesis was achieved through clinical case studies. The timeline for further evaluation was mapped out.

8.2 Future Directions

The research conducted for this thesis as well as the rich requirements emanating from the ever increasing amount of data available have provided a number of possible directions for future work. Some of the areas for future research are related closely to the contributions presented in this thesis. Initially the evaluation of the methodology and the prototype needs to be expanded through the deployment of two user studies. One user study will be aimed at the statistician and its aim will be to validate that the reasoning process and recommendation made by a statistician given the same research question and data is matched by the methodology proposed in this thesis. The second user study will be aimed at the end users and will assess how the use of such a prototype can reduce the time it takes clinicians to get from research question to a statistically robust conclusion, as well as eliciting general feedback on the usability of the prototype.

The feedback from both of these user studies would be incorporated into a more robust implementation. This would also include outputs specifically tailored to support the clinician, as well as offering an ontology-based input for research questions option. This will broaden the capability of this proposed methodology, as well as include a more comprehensive statistical knowledge base and access to more research questions.

The methodology proposed in this thesis could also be extended to the more general question of analysis approach automation beyond the clinical domain, this is also a future direction for my research. Domain areas rich in readily available data sets (such as geographical and government administrative data sets) would be suitable candidates for this.

A second area to address in future work is to address the relation between the selection of an approach to analyse data and the quality or trust in the data. This setting is of particular interest as it removes the assumption made in this thesis that the data is all in one table, and allows for the data required to answer a research question to make use of multiple data sources. This maps a more realistic scenario as it is rare that data is residing only in one location. The provenance and quality of each source of data, the degree of overlap in cases between each source and the clinician's perceived trust in the source will all affect how a research question is to be answered. This area bears some overlap with research on trust and data provenance in the computer science domain, as well as research on meta-analysis in statistics.

The process of recommending the most suitable analysis approach is illustrated in Figure 8.1 where the areas touched upon in this thesis are shaded in grey. The overall process shows the considerations that become relevant to the process once the assumption on relying only on one source of data is removed.

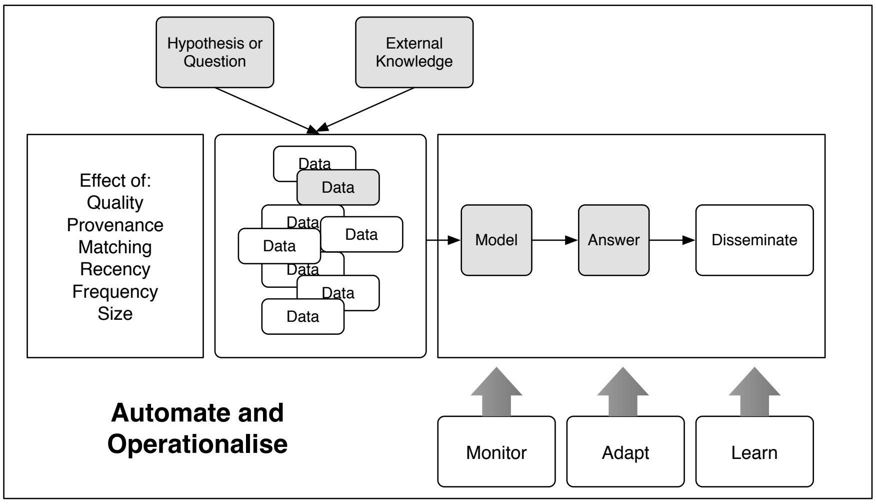


FIGURE 8.1: Future research areas vs research scope of thesis

Another relevant and related aspect to consider is the impact of missing data on the process illustrated in Figure 8.1 of statistical model selection, and the implications on the confidence in the resulting conclusions. Furthermore as the use of data from different sources is considered the patterns of missing data generated by collating the data is an area of interest. This can arise for example when a patient record is assembled by merging data from the Hospital IT systems and the data collected at departmental clinic surveys. In such cases this can result in some patients having complete records and others not having data across some or all of the data sources. This impacts the trust in the analysis process of the collated data. Furthermore the impact of these gaps in data will be dependant on the relative trust, recency and quality of each source of data.

The availability of multiple sources of data can also lead to the need to provide a hypothesis generation platform. This is another area for possible future research. This can be seen as an additional feature to enhance the proposed prototype based on the contributions of this thesis, as well as a standalone problem.

Although much of the research in this thesis has assumed that there is data, in future there is also a need to address the situation where the data needs to be collected. This can be achieved through automation of contents of the data to be collected, so that questionnaires can be devised to collect the data from the appropriate cohort and in adequate size to answer the research question. The latter either in isolation or as a complement to existing data sources.

8.3 Concluding Remarks

The availability of data and the accessibility of data continues to grow, and as such both the requirement for resources skilled to exploit data as well as the need for automation of data driven processes continues to grow. Data science is a relative new discipline related to the exploitation of data. Data science is now deemed one of the most appealing professions.

However the availability of skilled resources represents a bottleneck in the exploitation of data and as such methods that support self sufficiency of end users (not statisticians or data scientists) are required. Such methodologies would provide a range of levels of support to a wide range of data related tasks.

The research described in this thesis makes a contribution to the automation and support of the process of data analysis, and its emphasis is to support the non statistical end user by retaining transparency and involvement of the end user in the process.

In relating back to the cartoon illustrated in Figure 1.1 in chapter 1 of this thesis, I certainly hope that with increased emphasis on evidence based decision making, a plethora of tools and methodologies will be developed to make automated statistical testing of all 'gut feel' hypothesis a reality.

Appendix A

The Z Notation

This Appendix provides an introduction to Z notation using a simplified scenario related to the concepts required for statistical model selection. A glossary of the elements of Z notation can be found in [46] and sample specifications can be found in [49].

A.1 The Z specification language

Z is used to specify and to describe the behaviour of complex computer systems. A Z specification works by modelling the *states* that a system can take, the *operations* that cause changes in those states to take place and the *enquiries* that can discover information about those states.

A.1.1 Definitions

The Z notation makes use of typed set theory where all the possible elements or members of a set are considered to have something in common and are of the same type. For example, a set of people, or a set of numbers, but not a set that contains both types.

Within a specification *basic types* are chosen to be as widely encompassing as possible, with elements that are uniquely identifiable. For example a specification may refer to the set of all possible models available. This basic type would be [MODEL]. For each type it is also useful to introduce a listing of the identifiers of its elements. These could be for a type defined as [RESPONSE] for example: RESPONSE::=*yes* | *no*. Variables in the specifications must also be declared, so that the type of the value it refers to is stated. For example km:MODEL (Kaplan Meier is a model)

A.1.2 Initial state

Initially the system is described by defining the variables and any invariant properties relating those variables. Subsequently *operations* that *change* that state while maintaining the invariant properties. *Enquiries* can also be defined to obtain information about the system without changing its state.

A.1.3 Schemas

$[S]$

$a, b : \mathbb{N}$

$a < b$

A Z notation *specification* consists of a narrative text interspersed with formal descriptions written in Z notation. A graphical format called the *schema* was devised as a way of making a clear separation between the narrative text and the formal descriptions with a Z notation specification. An example schema is referred to by the name S and it declares two *variables* a and b. It also contains a *constraining predicate* which in this case states that a must be less than b. The top part of the schema contains the declarations, and the bottom part the predicate.

Schemas can be regarded as units and manipulated by various operators that are analogous to the logical operators. Z notation makes use of specific conventions on notation that I will now introduce.

Decoration: A schema name S' is the same as the schema S with its variables decorated with a prime. This signifies the value of a schema after some operation has been carried out. For example:

$[S']$

$a', b' : \mathbb{N}$

$a' < b'$

Schema **conjunction**: Two schemas can be joined by a *schema conjunction* operator, as a logical conjunction operator. The effect is to make a new schema with the declarations of the ctwo component schemas merged and their predicates conjoined. Given the example above:

$[T]$

$b, c : \mathbb{N}$

$b < c$

$$SandT == S \wedge T$$

is equivalent to:

$$\begin{array}{|l}
\hline
[SandT] \\
a, b, c : \mathbb{N} \\
\hline
a < b \\
b < c \\
\hline
\end{array}$$

Delta convention: The convention in Z notation is that the value of a variable before an operation is denoted by the undecorated name of the variable, and the value after an operation by the name decorated by a prime (') character, and is used in the *delta* naming convention.

$$\begin{array}{|l}
\hline
[\Delta S] \\
a, b : \mathbb{N} \\
a', b' : \mathbb{N} \\
\hline
a < b \\
a' < b' \\
\hline
\end{array}$$

The schema $[\Delta S]$ describes a state change.

Xi convention: The convention in Z notation is that a schema using Xi defines an operation in which the state does not change. A query of an existing knowledge base would employ this convention as there are no state changes involved. For example:

$$[\Xi S]$$

$$a, b : \mathbb{N}$$
$$a', b' : \mathbb{N}$$

$$a < b$$
$$a' < b'$$
$$a = a'$$
$$b = b'$$

A convention is used to denote the variables of a schema which specify operations. Finishing the variable's name with a question mark (?) indicates that the variable is an *input* to the schema. Finishing the variable's name with an exclamation mark (!) indicatess that the variable is an *output* of the schema.

Further conventions and symbols I will be making use of within the specification are:

- Domain anti-restriction $\lhd$: An object x is related to an object y by the relation$S \lhd R$ if and only if x is related to y by R and x is not a member of S.
- Range anti-restriction $\rhd$: An object x is related to an object y by the relation$R \rhd T$ if and only if x is related to y by R and y is not a member of T.

- Set comprehension: The members of the set $\{S \bullet E\}$ are the values taken by the expression E when the variables introduced by S take all possible values which make the property of S true.

A.1.4 Relations

The setting in which I am using Z notation involves more than one basic type. The proposed methodology's knowledge base consist of three basic types: MODEL, OBJECTIVE, ASSUMPTION. So a way of relating these sets to one another will be crucial. Relations within Z notation are based on the idea of a *cartesian product*, which is a pairing of values of two or more sets.

The relationship between the MODEL and OBJECTIVE in the context of the proposed methodology can be seen in Figure A.1. Such a set as illustrated in Figure A.1 can be declared:

$$achieves : \mathbb{P}(MODEL \times OBJECTIVE)$$

If a pair of values are related then their pair is an element in the relationship *achieves*. For example

$$(KM, time\_to\_event) \in achieves$$

A relation relates values of a set from a source to a target. The source involved in the relation is the *domain* or *dom* and the target is the *range* or *ran*. In this example

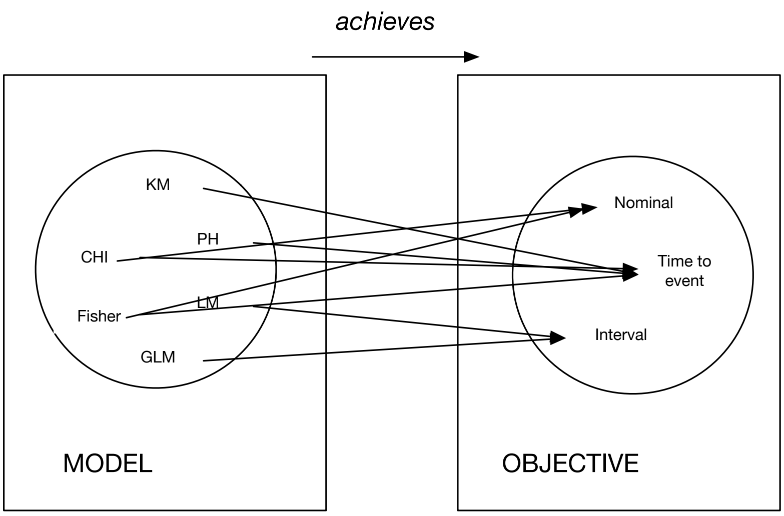


FIGURE A.1: SKB Relations: MODEL and OBJECTIVE

illustrated in Figure A.1:

$$\mathrm{dom}\, achieves = \{km, chi, ph, fisher, lm, glm\}$$

$$\mathrm{ran}\, achieves = \{time\_to\_event, nominal, interval\}$$

If we consider initially only the relation between MODEL and OBJECTIVE then this would be defined in Z as follows:

[MODEL] set of all models in scope

[OBJECTIVE] set of all objectives possible

The relationship between models and objectives is in the relation *achieves*:

$$achieves : MODEL \leftrightarrow OBJECTIVE$$

An operation to ascertain whether model $m?$ achieves objective $o?$ would generate a variable REPLY:

$$REPLY ::= yes \mid no$$

This would be used as follows:

$$
\begin{array}{|l}
\hline
[Enquire] \\
\Xi achieves \\
m? : MODEL \\
o? : OBJECTIVE \\
rep! : REPLY \\
\hline
(m? \mapsto o? \in achieve \land rep! = yes) \\
\lor \\
(m? \mapsto o? \notin achieve \land rep! = no) \\
\hline
\end{array}
$$

The relation between MODEL and OBJECTIVE will also rely on schemas to populate it and to maintain it. The proposed knowledge base relies on an additional relation, between the models and the critical assumptions. This will also need to be represented in Z and in order to so so the relations are joined together by an operation called *composition*.

An example of the relation that needs to be modelled is in Figure A.2.

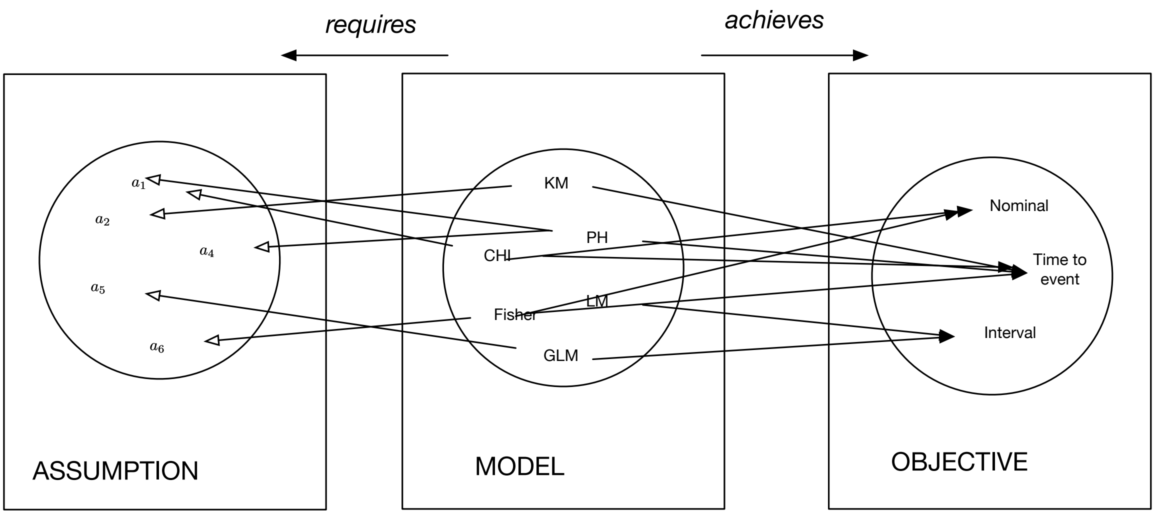


FIGURE A.2: SKB Relations: MODEL × OBJECTIVE × ASSUMPTION

The relation formed by the relation *require*, then the relation *achieves* is a *forward composition* or *reqiores* with *achieves*:

$$require : ASSUMPTION \leftrightarrow MODEL$$
$$achieve : MODEL \leftrightarrow OBJECTIVE$$
$$require;\ achieve : ASSUMPTION \leftrightarrow OBJECTIVE$$

Appendix B

Statistical Knowledge Base contents

This appendix includes an overview of the models within the SKB relevant to the case studies related to this thesis. The objective is directly mapped to the type of the dependent variable within the research question or hypothesis. Three objectives are currently covered in this document (time to event, interval and nominal).

B.1 Time to event (S)

Models:

1. KM (Kaplan Meier) [47] $m_{s1} = km$
2. PH (Cox Proportional Hazards) [24] $m_{s2} = ph$
3. Weibull [86] $m_{s3} = w$

Note: that in a more comprehensive statistical knowledge base models such as competing risk models and survival forests would be added.

Alternative objective

The alternative method of analysing time to event target variable is to transform the time to event column into a binary one by selecting an appropriate time cut off. The binary variable can then be analysed using $O_{alt} = o_n$.

- If t is the time to event variable then t_b
- T is the relevant cut off time. Typically this would be 3 years or 5 years
- A new target variable is calculated as follows: $t' = 1 \mid t \geq T$ else $t' = 0$
- t' will be a binary variable

Assumptions:

1. Non informative censoring a_1
2. Testing for proportional Hazards a_2
3. Testing for Weibull distribution a_3

Mapping of assumptions critical to models

- $m_{s1} = km \mapsto a_1$
- $m_{s2} = ph \mapsto a_1, m_{s2} = ph \mapsto a_2$
- $m_{s3} = w \mapsto a_1, m_{s3} = w \mapsto a_3$

R code to test assumptions

In order to test assumption a_2:

```
#Proportional Hazards model
ovarian.ph<-coxph(Surv(ovarian$futime, ovarian$fustat==1)~ovarian$rx,data=ovarian)
summary(ovarian.ph)


#testing the proportional hazards assumption -
# if the p value resulting in this is <0.05
#then the hazards are not proportional and the assumption does not hold.
ovarian.zph<-cox.zph(ovarian.ph,transform = 'log')
ovarian.zph
```

In order to test assumption a_3

```
#Testing the weibull assumption:
#The estimated log log lines in the graph produced should be roughly
#straight if the Weibull model is appropriate.
m=1
n=ovarian.survfit$strata[1]
temp<-ovarian.survfit$time[m:n]
cloglog=log(-log(ovarian.survfit$surv[m:n]))
plot(log(temp), cloglog, type ="s")
m=n+1
n=n+ovarian.survfit$strata[2]
temp=ovarian.survfit$time[m:n]
cloglog=log(-log(ovarian.survfit$surv[m:n]))
lines(log(temp),cloglog,type="s",col=2)
```

context domain	model m	Performance measure p
$cd1_1$ censoring absent	m_{s1} KM	p_1 = unaffected
	m_{s2} PH	p_1 = unaffected
	m_{s3} Weibull	p_1 = unaffected
	m_{b1} χ^2	p_1 = unaffected
	m_{b2} $Fisher's$	p_1 = unaffected
$cd1_2$ censoring light	m_{s1} KM	p_2 = mildly affected
	m_{s2} PH	p_1 = unaffected
	m_{s3} Weibull	p_1 = unaffected
	m_{b1} χ^2	p_3 = strongly affected
	m_{b2} $Fisher's$	p_3 =strongly affected
$cd1_3$ censoring heavy	m_{s1} KM	p_3 = strongly affected
	m_{s2} PH	p_1 =unaffected
	m_{s3} Weibull	p_3 = strongly affected
	m_{b1} χ^2	p_3 = strongly affected
	m_{b2} $Fisher's$	p_3 = strongly affected

Table B.1: $cd1_1$, $cd1_2$ and $cd1_3$ performance function mapping

context domain	model m	Performance measure p
$cd2_1$predict	m_{s1} KM	p_3 avoid
	m_{s2} PH	p_1 suitable
	m_{s3} Weibull	p_1 suitable
	m_{b1} χ^2	p_3 avoid
	m_{b2} $Fisher's$	p_3 avoid
$cd2_2$ explain	m_{s1} KM	p_1 suitable
	m_{s2} PH	p_1 suitable
	m_{s3} Weibull	p_1 suitable
	m_{b1} χ^2	p_2 neutral
	m_{b2} $Fisher's$	p_2 neutral

Table B.2: $cd2_1$ and $cd2_2$ performance function mapping for model intent

Context domains:

- $CD1_1, CD1_2, CD1_3$ Censoring levels *absent, light, heavy*
- $CD2_1, CD2_2$ Model Intent *predict, explain*

In order to identify if $CD1_1$, $CD1_2$ or $CD1_3$ is present:

```
#Proportion of censored cases - a patient is censored if fustat=0
```

```
#(the event has not been observed yet so we know they
# were followed for at leas the value in futime )
#no censoring = 0, mild censoring < 0.7, heavy censoring >= 0.7

censoring.prop<-(nrow(ovarian)-sum(ovarian$fustat))/nrow(ovarian)
```

CD_2 will depend on the declared analysis intent of the clinician.

The order of performance measures relevant to these CD is:

- cd_1: p_1 unaffected $\succ$ p_2 mildly affected $\succ$ p_3 strongly affected
- cd_2: p_1 suitable $\succ$ p_2 neutral $\succ$ p_3 avoid

B.2 Interval (I)

Models:

1. anova [25] $m_{i1} = anova$
2. t-test [76] $m_{i2} = t$
3. Welch [87] $m_{i3} = welch$

Assumptions:

1. Is the independent variable normally distributed? a_4
2. Independent variable (or covariate of interest) is nominal? a_5
3. independent variable (or covariate of interest) is binary? a_6
4. Is the variance equal within each level of the target variable? (Homoscedasticy) a_7

Mapping of assumptions critical to models

- $m_{i1} = anova \mapsto a_4, m_{i1} = anova \mapsto a_7$
- $m_{i2} = t \mapsto a_4, m_{i2} = t \mapsto a_6, m_{i2} = t \mapsto a_7$
- $m_{i3} = welch \mapsto a_4, m_{i3} = welch \mapsto a_6$

Context Domains

- Missing data
- Model intent

B.3 Categorical (N)

Models:

1. chi squared [64] $m_{b1} = \chi^2$
2. Fisher's exact [34] $m_{b2} = Fisher$
3. Logistic regression $m_{b3} = lr$
4. Decision Tree $m_{b4} = dt$
5. Neural Network - $m_{b5} = nn$
6. Linear Discriminant Analysis - $m_{b6} = lda$

Assumptions:

1. Is the dependent variable binary? a_8
2. Are there more than 500 cases? a_9
3. Are there more than 5 cases in one cell? a_{10}
4. Is there more than one covariate of interest? a_{11}
5. Are the independent variable multivariate normally distributed? a_{12}

Mapping of assumptions critical to models

- $m_{b1} = \chi^2 \mapsto a_{10}, m_{b1} = \chi^2 \mapsto \neg\, a_{11}$
- $m_{b2} = Fisher \mapsto \neg\, a_{11}$
- $m_{b3} = lr \mapsto a_{11}, m_{b3} = lr \mapsto a_8$

Context Domain	Model	Performance measure
$cd2_1$ predict	$m_{b1}\chi^2$	p_3 avoid
	m_{b2} Fisher's	p_3 avoid
	m_{b3} lr	$p_1$suitable
	m_{b4} dt	p_2 neutral
	m_{b5} nn	p_1 suitable
	m_{b6} lda	p_1 suitable
$cd2_2$ explain	$m_{b1}\chi^2$	p_1 suitable
	m_{b2} Fisher's	p_1 suitable
	m_{b3} lr	p_1 suitable
	m_{b4} dt	p_1 suitable
	m_{b5} nn	p_3 avoid
	m_{b6} lda	p_2 neutral

Table B.3: Sample performance function for model intent for objective o_n

Context Domain	Model	Performance measure
$cd3$ missing data (*)	$m_{b1}\chi^2$	p_1 unaffected
	m_{b2} Fisher's	p_1 unaffected
	m_{b3} lr	p_2 affected
	m_{b4} dt	p_1 unaffected
	m_{b5} nn	p_2 affected
	m_{b6} lda	p_1 unaffected

Table B.4: Sample performance function for missing data for objective o_n

- $m_{b4} = dt \mapsto a_9, m_{b4} = dt \mapsto a_{11}$
- $m_{b6} = lda \mapsto a_8, m_{b6} = lda \mapsto a_{12}$

Context domains

- CD_1 Model intent
- CD_2 Missing data

(*) Note that the percentage of missing data can be assessed in a number of ways. Furthermore it is also possible that different patterns of missing data can negate the use of a model or favour it. For simplicity an arbitrary percentage of records with at least one of the explanatory variables missing will be used. In future this can be refined further.

Bibliography

[1] Víctor Abraira, Alfonso Muriel, José I Emparanza, José I Pijoan, Ana Royuela, María Nieves Plana, Alejandra Cano, Iratxe Urreta, and Javier Zamora. Reporting quality of survival analyses in medical journals still needs improvement. a minimal requirements proposal. *Journal of Clinical Epidemiology*, 66(12):1340–1346, Dec 2013.

[2] Constantin Aliferis, Evelyn Chao, and Gregory F Cooper. Data explorer: a prototype expert system for statistical analysis. In *Proceedings of the Annual Symposium on Computer Application in Medical Care*, pages 389– 393. American Medical Informatics Association, 1993.

[3] Leila Amgoud and Claudette Cayrol. Inferring from inconsistency in preference-based argumentation frameworks. *Journal of Automated Reasoning*, 29(2):125–169, 2002.

[4] Leila Amgoud and Claudette Cayrol. A reasoning model based on the production of acceptable arguments. *Annals of Mathematics and Artificial Intelligence*, 34(1-3):197–215, 2002.

[5] Leila Amgoud, Simon Parsons, and Laurent Perrussel. An argumentation framework based on contextual preferences. In *Formal and Applied Practical Reasoning (FAPR 2000)*, pages 59–67, 2000.

[6] Leila Amgoud and Srdjan Vesic. Repairing preference-based argumentation frameworks. In *Proceedings of the 21st International Joint Conference on Artificial Intelligence,*, pages 665–670, 2009.

[7] Leila Amgoud and Srdjan Vesic. Rich preference-based argumentation frameworks. *International Journal of Approximate Reasoning*, 55(2):585–606, 2014.

[8] Chris Anderson. The end of theory: the data deluge makes the scientific method obsolete. wired magazine 16.07, 2008.

[9] Katie Atkinson. What should we do?: Computational representation of persuasive argument in practical reasoning. *PhD Thesis, University of Liverpool*, 2005.

[10] Katie Atkinson and Trevor Bench-Capon. *Abstract Argumentation Scheme Frameworks*, pages 220–234. Springer Berlin Heidelberg, Berlin, Heidelberg, 2008.

[11] Katie Atkinson, Trevor Bench-Capon, and Peter McBurney. Arguing about cases as practical reasoning. In *Proceedings of the 10th international conference on Artificial intelligence and law*, pages 35–44. ACM, 2005.

[12] Katie Atkinson, Trevor Bench-Capon, and Peter McBurney. A dialogue game protocol for multi-agent argument over proposals for action. *Autonomous Agents and Multi-Agent Systems*, 11(2):153–171, 2005.

[13] Katie Atkinson, Trevor J. M. Bench-Capon, and Sanjay Modgil. Argumentation for decision support. In Stéphane Bressan, Josef Küng, and Roland Wagner, editors, *DEXA*, volume 4080 of *Lecture Notes in Computer Science*, pages 822–831. Springer, 2006.

[14] Pietro Baroni, Martin Caminada, and Massimiliano Giacomin. Review: An introduction to argumentation semantics. *The Knowledge Engineering Review*, 26(4):365–410, December 2011.

[15] Trevor J. M. Bench-Capon and Paul E. Dunne. Argumentation in artificial intelligence. *Artificial Intelligence*, 171(10-15):619–641, 2007.

[16] Trevor JM Bench-Capon. Persuasion in practical argument using value-based argumentation frameworks. *Journal of Logic and Computation*, 13(3):429–448, 2003.

[17] Philippe Besnard and Anthony Hunter. *Elements of Argumentation.* The MIT Press, 2008.

[18] George EP Box. Sampling and bayes' inference in scientific modelling and robustness. *Journal of the Royal Statistical Society. Series A (General)*, pages 383–430, 1980.

[19] Martin Caminada. A gentle introduction to argumentation semantics. *Lecture material*, 2008.

[20] Martin Caminada and Leila Amgoud. On the evaluation of argumentation formalisms. *Artificial Intelligence*, 171(5–6):286 – 310, 2007.

[21] Martin Caminada and Yining Wu. On the limitations of abstract argumentation. In *Benelux Conference on Artificial Intelligence (BNAIC 2011).* 2011.

[22] Sylvie Coste-Marquis, Jérôme Lang, Paolo Liberatore, and Pierre Marquis. Expressive power and succinctness of propositional languages for preference representation. volume 4, pages 203–212, 2004.

[23] AS Coulson, DW Glasspool, J Fox, and J Emery. Rags: A novel approach to computerized genetic risk assessment and decision support from pedigrees. *Methods Of Information in Medicine*, 40(4):315–322, 2001.

[24] D. R. Cox. Regression models and life-tables. *Journal of the Royal Statistics Society, Series B*, 34:187–220, 1972.

[25] D. R. Cox and Nancy Reid. *The theory of the design of experiments*. CRC Press, 2000.

[26] Madalina Croitoru, Jerome Fortin, and Oren Nir. Arguing with preferences in ecobiocap. In *COMMA: Computational Models of Argument*, volume Frontiers in Artificial Intelligence and Applications of *Computational Models of Argument*, pages 51–58, Vienne, Austria, Sept 2012. IOS Press.

[27] Kristijonas Čyras. Argumentation-based reasoning with preferences. In *Highlights of Practical Applications of Scalable Multi-Agent Systems. The PAAMS Collection*, pages 199–210. Springer, 2016.

[28] Maria Dami and Alfred Vella. Expert systems in statistics. *SPOUDAI Journal of Economics and Business*, 38(3):390–402, 1988.

[29] Mark d'Inverno, Michael Luck, Michael Georgeff, David Kinny, and Michael Wooldridge. The dmars architecture: A specification of the distributed multi-agent reasoning system. *Autonomous Agents and Multi-Agent Systems*, 9(1-2):5–53, 2004.

[30] Carmel Domshlak, Eyke Hüllermeier, Souhila Kaci, and Henri Prade. Preferences in ai: An overview. *Artificial Intelligence*, 175(7–8):1037 – 1052, 2011. Representing, Processing, and Learning Preferences: Theoretical and Practical Challenges.

[31] Phan Minh Dung. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial Intelligence*, 77(2):321–358, 1995.

[32] Phan Minh Dung, Robert A Kowalski, and Francesca Toni. *Assumption-based argumentation*, pages 199–218. Springer US, Boston, MA, 2009.

[33] John H Edmonson, Thomas Richard Fleming, DG Decker, GD Malkasian, EO Jorgensen, JA Jefferies, MJ Webb, and LK Kvols. Different chemotherapeutic sensitivities and host factors affecting prognosis in advanced ovarian carcinoma versus minimal residual disease. *Cancer treatment reports*, 63(2):241–247, 1979.

[34] Ronald A Fisher. On the interpretation of χ 2 from contingency tables, and the calculation of p. *Journal of the Royal Statistical Society*, 85(1):87–94, 1922.

[35] Ronald A Fisher. Presidential address. *Sankhyā: The Indian Journal of Statistics*, pages 14–17, 1938.

[36] John Fox, David Glasspool, and Jonathan Bury. *Quantitative and Qualitative Approaches to Reasoning under Uncertainty in Medical Decision Making*, volume 2101 of *Lecture Notes in Computer Science*, pages 272–282. Springer Berlin Heidelberg, 2001.

[37] John Fox, David Glasspool, Dan Grecu, Sanjay Modgil, Matthew South, and Vivek Patkar. Argumentation-based inference and decision making–a medical perspective. *IEEE Intelligent Systems*, 22(6):34–41, 2007.

[38] John Fox, Nicky Johns, and Ali Rahmanzadeh. Disseminating medical knowledge: the proforma approach. *Artificial Intelligence in Medicine*, 14(1):157–182, 1998.

[39] David Glasspool, John Fox, Ayelet Oettinger, and James Smith-Spark. Argumentation in decision support for medical care planning for patients and clinicians. In *AAAI Spring Symposium: Argumentation for Consumers of Healthcare*, pages 58–63. AAAI, 2006.

[40] Nikos Gorogiannis, Anthony Hunter, and Matthew Williams. An argument-based approach to reasoning with clinical knowledge. *International Journal of Approximate Reasoning*, 51(1):1–22, 2009.

[41] María Adela Grando, Laura Moss, David Glasspool, Derek H. Sleeman, Malcolm Sim, Charlotte J. Gilhooly, and John Kinsella. Argumentation-logic for explaining anomalous patient responses to treatments. In Mor Peleg, Nada Lavrac, and Carlo Combi, editors, *Artificial Intelligence in Medicine*, volume 6747 of *Lecture Notes in Computer Science*, pages 35–44. Springer, 2011.

[42] María Adela Grando, Laura Moss, Derek H. Sleeman, and John Kinsella. Argumentation-logic for creating and explaining medical hypotheses. *Artificial Intelligence in Medicine*, 58(1):1–13, 2013.

[43] D. J. Hand. Statistical expert systems: design. *The Statistician*, pages 351–369, 1984.

[44] Anthony Hunter and Matthew Williams. Using clinical preferences in argumentation about evidence from clinical trials. In *Proceedings of the 1st ACM International Health Informatics Symposium*, IHI '10, pages 118–127, New York, NY, USA, 2010. ACM.

[45] Anthony Hunter and Matthew Williams. Aggregating evidence about the positive and negative effects of treatments. *Artificial Intelligence in Medicine*, 56(3):173–190, 2012.

[46] J Jacky. *The Way of Z: Practical Programming with Formal Methods.* Cambridge University Press, 1997.

[47] E. L. Kaplan and Paul Meier. Nonparametric Estimation from Incomplete Observations. *Journal of the American Statistical Association*, 53(282):457–481, 1958.

[48] Paul Krause, Simon Ambler, Morten Elvang-Goransson, and John Fox. A logic of argumentation for reasoning under uncertainty. *Computational Intelligence*, 11(1):113–131, 1995.

[49] D Lightfoot. *Formal Specification using Z.* Palgrave, 2001.

[50] Helena Lindgren. Towards using argumentation schemes and critical questions for supporting diagnostic reasoning in the dementia domain. *Proceedings of Computational Models of Natural Arguments (CMNA'09), Pasadena, CA*, pages 10–14, 2009.

[51] James Robert Lloyd, David K. Duvenaud, Roger B. Grosse, Joshua B. Tenenbaum, and Zoubin Ghahramani. Automatic construction and natural-language description of nonparametric regression models. *CoRR*, abs/1402.4304, 2014.

[52] Michael Luck and Mark d'Inverno. A conceptual framework for agent definition and development. *The Computer Journal*, 44(1):1–20, 2001.

[53] N Mantel. Evaluation of survival data and two new rank order statistics arising in its consideration, cancer chemotherapy rep., 50, 163–170. *Google Scholar*, 1967.

[54] Peter McBurney. What are models for? In *Proceedings of the 9th European Conference on Multi-Agent Systems*, EUMAS'11, pages 175–188, Berlin, Heidelberg, 2012. Springer-Verlag.

[55] Peter McBurney, David Hitchcock, and Simon Parsons. The eightfold way of deliberation dialogue. *International Journal of Intelligent Systems*, 22(1):95–132, 2007.

[56] Tim Miller and Peter McBurney. Multi-agent system specification using tcoz. In *German Conference on Multiagent System Technologies*, pages 216–221. Springer, 2005.

[57] Andrei Mocanu, Xiuyi Fan, Francesca Toni, Matthew Williams, and Jiarong Chen. Recomedic: recommending medical literature through argumentation. In *4th International Workshop on Combinations of Intelligent Methods and Applications (CIMA 2014). Limassol, Cyprus*, pages 10–11, 2014.

[58] Sanjay Modgil. Reasoning about preferences in argumentation frameworks. *Artificial Intelligence*, 173(9–10):901 – 934, 2009.

[59] Sanjay Modgil and Henry Prakken. Revisiting preferences and argumentation. In *IJCAI Proceedings-International Joint Conference on Artificial Intelligence*, volume 22, page 1021, 2011.

[60] Sanjay Modgil and Henry Prakken. The ASPIC+ framework for structured argumentation: a tutorial. *Argument & Computation*, 5(1):31–62, 2014.

[61] Simon Parsons, Elizabeth Sklar, Jordan Salvit, Holly Wall, and Zimi Li. Argtrust: decision making with information from sources of varying trustworthiness. In *Proceedings of the 2013 international conference on Autonomous agents and multi-agent systems*, pages 1395–1396. International Foundation for Autonomous Agents and Multiagent Systems, 2013.

[62] V Patel, M. McGurk, C. Sproat, J. Kwok, Y Gadiwalla, and I. Sassoon. Prophylactic use of pentoxifylline and tocopherol in irradiated head and neck oncology patients requiring dental extractions. *The British Journal of Oral and Maxillofacial Surgery*, 54(3):342–345, Jan 2015.

[63] Vinod Patel, Yusuf Gadiwalla, Isabel Sassoon, Chris Sproat, Jerry Kwok, and Mark McGurk. Use of pentoxifylline and tocopherol in the management of osteoradionecrosis. *British Journal of Oral and Maxillofacial Surgery*, 54(5):547–550, March 2016.

[64] Karl Pearson. X. on the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 50(302):157–175, 1900.

[65] Henry Prakken. An abstract framework for argumentation with structured arguments. *Argument and Computation*, 1(2):93–124, 2010.

[66] C Radhakrishna Rao. *Linear statistical inference and its applications*. John Wiley & Sons, 2009.

[67] Chris Reed, Douglas Walton, and Fabrizio Macagno. Argument diagramming in logic, law and artificial intelligence. *Knowledge Engineering Review*, 22(1):87–109, 2007.

[68] Isabel Sassoon, Jeroen Keppens, and Peter McBurney. Towards argumentation for statistical model selection. In *Fifth International Conference on Computational Models of Argument*, pages 67–74. IOS Press, 2014.

[69] Isabel Sassoon, Jeroen Keppens, and Peter McBurney. Preferences in argumentation for statistical model selection. In *Sixth International Conference on Computational Models of Argument*, pages 53–60. IOS Press, 2016.

[70] Clare Schilling, Sandro J. Stoeckli, Stephan K. Haerle, Martina A. Broglie, Gerhard F. Huber, Jens Ahm Sorensen, Vivi Bakholdt, Annelise Krogdahl, Christian von Buchwald, Anders Bilde, Lars R. Sebbesen, Edward Odell, Benjamin Gurney, Michael O'Doherty, Remco de Bree, Elisabeth Bloemena, Geke B. Flach, Pedro M. Villarreal, Manuel Florentino Fresno Forcelledo, Luis Manuel Junquera Gutiarrez, Julio Alvarez Amazaga, Luis Barbier, Joseba SantamarÃ a Zuazua, Augusto Moreira, Manuel Jacome, Maurizio Giovanni Vigili, Siavash Rahimi, Girolamo Tartaglione, Georges Lawson, Marie-Cecile Nollevaux, Cesare Grandi, Davide Donner, Emma Bragantini, Didier Dequanter, Philippe Lothaire, Tito Poli, Enrico M. Silini, Erinco Sesenna, Giles Dolivet, Romina Mastronicola, Agnes Leroux, Isabel Sassoon, Philip Sloan, and Mark McGurk. Sentinel european node trial

(sent): 3-year results of sentinel node biopsy in oral cancer. *European Journal of Cancer*, 51(18):2777 – 2784, 2015.

[71] Floarea Serban, Joaquin Vanschoren, Jörg-Uwe Kietz, and Abraham Bernstein. A survey of intelligent assistants for data analysis. *ACM Computing Surveys (CSUR)*, 45(3):31, 2013.

[72] Galit Shmueli. To explain or to predict? *Statistical Science*, pages 289–310, 2010.

[73] Elizabeth Sklar, Simon Parsons, Zimi Li, Jordan Salvit, Senni Perumal, Holly Wall, and Jennifer Mangels. Evaluation of a trust-modulated argumentation-based interactive decision-making tool. *Autonomous Agents and Multi-Agent Systems*, 30(1):136–173, 2016.

[74] J Michael Spivey and JR Abrial. *The Z Notation*. Prentice Hall Hemel Hempstead, 1992.

[75] Robert Stratton. *Automated Theory Selection using Agent Based Models*. PhD thesis, King's College London, 2015.

[76] Student. The probable error of a mean. *Biometrika*, pages 1–25, 1908.

[77] Girolamo Tartaglione, Sandro J Stoeckli, Remco De Bree, Clare Schilling, Geke B Flach, Vivi Bakholdt, Jens Ahm Sorensen, Anders Bilde, Christian Von Buchwald, Georges Lawson, et al. Sentinel node in oral cancer: The nuclear medicine aspects. a survey from the sentinel european node trial. *Clinical nuclear medicine*, 41(7):534–542, 2016.

[78] David Tighe, Isabel Sassoon, A. Kwok, and Mark McGurk. Is benchmarking possible in audit of early outcomes after operations for head and neck cancer? *British Journal of Oral and Maxillofacial Surgery*, 52(10):913–921, Sept 2014.

[79] David Tighe, Alan J Thomas, Robin Kinsman, Isabel Sassoon, and Mark McGurk. Development of a benchmarking tool for audit of early outcomes after surgery for head and neck squamous cell carcinoma. *British Journal of Oral and Maxillofacial Surgery*, 54(10):e177, 2016.

[80] David F Tighe, Alan J Thomas, Isabel Sassoon, Robin Kinsman, and Mark McGurk. Developing a risk stratification tool for audit of outcome after surgery for head and neck squamous cell carcinoma. *Head & Neck*, 2017.

[81] Pancho Tolchinsky, Ulises Cortés, Sanjay Modgil, Francisco Caballero, and Antonio López-Navidad. Increasing human-organ transplant availability: Argumentation-based agent deliberation. *IEEE Intelligent Systems*, 21(6):30–37, 2006.

[82] Pancho Tolchinsky, Sanjay Modgil, Katie Atkinson, Peter McBurney, and Ulises Cortés. Deliberation dialogues for reasoning about safety critical actions. *Autonomous Agents and Multi-Agent Systems*, 25(2):209–259, 2012.

[83] S.E. Toulmin. *The Uses of Argument*. Cambridge University Press, 1958.

[84] Douglas Walton. *Argumentation Theory: A Very Short Introduction*, pages 1–22. Springer US, 2009.

[85] Douglas Walton, Christopher Reed, and Fabrizio Macagno. *Argumentation Schemes*. Cambridge University Press, 2008.

[86] Waloddi Weibull. Wide applicability. *Journal of applied mechanics*, 1951.

[87] Bernard L Welch. The generalization of student's' problem when several different population variances are involved. *Biometrika*, 34(1/2):28–35, 1947.

[88] Richard Whately. *Elements of Logic*. Scholars' Facsimiles & Reprints, 1827.

[89] Sebastian Zillessen. Small Data Analyst, MSc. Thesis, King's College London. Aug 2016.

Index